

Fear and loathing at Los Alamos

The laboratory that gave birth to the nuclear bomb is caught in an unsettling downward spiral of weak leadership and dwindling staff morale.

The plagues afflicting Los Alamos National Laboratory in New Mexico are beginning to reach Biblical proportions. Since 1999, when the lab became the target of a so-far inconclusive probe into the apparent leaking of warhead design details to China, it has endured multiple security crises, forest fires and congressional investigations into the alleged theft of computer equipment by employees. This summer, more computer disks containing classified information have disappeared, causing Peter Nanos, the lab's director, to place 19 employees on administrative leave and suspend all research activities indefinitely (see *Nature* **430**, 387; 2004).

The suspension holds up not just work on nuclear weapons, but also an array of non-classified research ranging from genetics to computer science. Adding insult to injury, media accounts of the suspension portray the Los Alamos scientific staff as arrogant and disdainful of the bureaucratic necessities of handling classified data.

Yet the latest security crisis should not be taken at face value. The contract for running the lab, held by the University of California since 1943, is up for competitive tender next October. Two universities from the politically well-connected state of Texas — the University of Texas and Texas A&M — have expressed an interest in winning the contract. This might explain why Congressman Joe Barton (Republican, Texas) saw fit to visit Los Alamos as soon as the disks went missing and then called for an FBI investigation into their disappearance.

There is a growing impression that Los Alamos is not being treated fairly. It has emerged, for example, that it was not the only nuclear-weapons lab to lose some classified material this summer: in late June, Sandia National Laboratories in nearby Albuquerque misplaced a disk containing such data for three weeks. No alarm bells were sounded in Washington, however, and there have been no calls to re-examine Lockheed Martin's contract to manage Sandia.

The lock-down at Los Alamos has, according to senior scientists there, had a profoundly negative impact on laboratory morale, which was already beaten down by the prospect of staff losing their valued academic affiliations with the University of California if it loses the contract. And Nanos has blasted his own staff for what he termed a "cowboy culture" at the laboratory; the tone of his public statements suggests a frightening gulf between the leader and the led.

Energy secretary Spencer Abraham, who heads the government agency that oversees Los Alamos, has also denounced the lab's employees. Linton Brooks, head of the National Nuclear Security Administration, which was set up in response to the 1999 scandal to improve management at the nuclear-weapons laboratories, has said and done nothing. And even Senator Pete Domenici (Republican, New Mexico), the most stalwart defender of Los Alamos in Congress, despairs of its future: "Today, in Washington, Los Alamos' reputation as a crown jewel of science is being eclipsed by a reputation as being both dysfunctional and untouchable," he declared in an open letter to laboratory employees on 22 July.

Any security lapse at a nuclear-weapons laboratory must be taken seriously. At Los Alamos, classified information is currently stored in more than 2,000 safes spread across its sprawling complex; it probably should be centralized along lines already used at the rival Lawrence Livermore National Laboratory in California. But without strong, credible leadership and the full cooperation of researchers, such reforms cannot be properly implemented.

It is intolerable that a national resource as important as Los Alamos should be allowed to languish for years as a political football, losing people and prestige with every bounce. It is time that Abraham, Domenici, Nanos and others sat down to decide how to give people at Los Alamos some degree of stability until the new contract is let. ■

An opportunity lost

No one is taking responsibility for tracking the pathological aftermath of inconclusive trials of an Alzheimer's vaccine.

Valuable data on patients' responses to a candidate vaccine for treating Alzheimer's disease may never be collected, now that the San Francisco arm of Elan has abandoned its trial (see page 715). The case highlights a quandary that can occur when scientifically interesting trials run out of commercial promise. The issue doesn't arise with most trials of drug therapies or medical devices, as these have no further biological influence when the trial is over. But the vaccine trial is different: long-term follow-up might shed light on the pathological consequences of Alzheimer's vaccines.

In January 2002, Elan stopped its trial of a vaccine against the amyloid- β protein, a major constituent of the destructive brain plaques that occur in Alzheimer's, after some patients suffered brain inflammation. But many of the participants developed antibodies to amyloid- β , and these are still biologically active.

Elan reported the results of the trial, taking into account clinical data gathered until December 2002. But only one of the 28 medical centres that provided patients for the trial continues to study its

patients systematically — and preliminary results suggest that the course of progress changes in the second year of immunization.

A group of patients with antibodies to amyloid- β is a unique resource. But who should organize, and pay for, the monitoring of the group? Elan won't do it and the US Food and Drug Administration says it has no jurisdiction over what happens when a trial ends early. It will intervene only if a company is close to licensing a product, as it did, for example, to insist on the long-term monitoring of patients in gene-therapy trials. It says that placing the burden of responsibility on companies such as Elan would be a disincentive to innovation. Research funding agencies, such as the US National Institutes of Health, are also disinclined to finish off the trials of others.

Funding agencies should be more sympathetic to grant applications from researchers in participating centres who are willing to take on the onerous task of monitoring the progress of patients scattered around the world. Such work may not be particularly innovative, but it could yield important health benefits and so deserves support. ■





In circulation
Blood transfusions
could perpetuate
vCJD epidemic
p712



Fuel for thought
Trouble on the
horizon for NASA's
mission to Pluto
p713



Cold comfort
Polar scientists voice
fears over Antarctic
bioprospectors
p714



Missed chance
Doctors say valuable
data on Alzheimer's
vaccine were lost
p715

Victims hit out at university over handling of harassment cases

Erika Check

Officials affiliated to the largest research university in Canada have been accused of mis-handling complaints of sexual harassment.

Two female students, who undertook research at teaching hospitals affiliated to the University of Toronto, say that administrators conducted flawed investigations, which have damaged their scientific careers. Officials at the University Health Network (UHN), which runs research at the hospitals, admit that harassment took place but say appropriate steps were taken to tackle it — a claim that may soon be examined by human-rights officials.

The most serious allegations have been made by Gwen Schwartz, who was doing doctoral research on spinal injuries at Toronto Western Hospital until last year. Schwartz says that she began a sexual relationship with her supervisor, Michael Fehlings, in December 2000. When she attempted to end this relationship the following month, Schwartz alleges that Fehlings began sexually harassing her and that the harassment continued until she complained to UHN officials in March 2003.

The complaint eventually led to a settlement signed last December between Schwartz, Fehlings, the university and the hospital. Schwartz was offered some Can\$200,000 (US\$152,000) to cover legal fees and damages, and agreed to withdraw her complaint and not to file another in the future.

But Schwartz is unhappy with both the investigation and the settlement. She claims that she was removed from her lab during the inquiry, losing access to her thesis data. Schwartz also regrets signing the settlement, saying she only did so because she feared she could not pay legal fees. She adds that Fehlings has still not given her access to all of her data and she is currently refusing to accept the money due to her under the settlement.

"This has completely stalled my academic career, jeopardized my education and and threatened my economic independence," Schwartz says.

Fehlings told *Nature* that he "vehemently denies Schwartz's allegations", but says the settlement he has reached with her prevents



Toronto student Loralyn Benoit claims that her complaint about sexual harassment was handled inappropriately by the University Health Network, which is headed by Tom Closson (inset).

him from discussing details of the case.

The second graduate student, immunologist Loralyn Benoit, made accusations of sexual harassment in February 2003 against Claude Cantin, the manager running her lab in Toronto's Princess Margaret Hospital. Following a UHN investigation that ended in November 2003, the hospital concluded that harassment had occurred.

Benoit is now writing up her thesis, but she too is angry about the way her complaint was handled. She alleges that senior university officials asked her inappropriate questions of a sexual nature during the investigation. Officials at UHN will not reveal what, if any, action was taken as a result of their inquiry.

Benoit adds that she pursued her complaint with the help of a lawyer, but was advised to end negotiations when the university asked her to sign an agreement that included a confidentiality clause.

Schwartz and Benoit say that their stories raise grave concerns about the University of Toronto's ability to protect students doing research at university-affiliated hospitals. In January, Benoit asked the Ontario Human Rights Commission to examine the way the

hospital dealt with her case. The commission has agreed to do so and is assessing a similar application from Schwartz made in June.

Senior officials within the UHN have said in the Canadian press that they believe sexual harassment occurred in both cases, yet deny that their investigations were flawed.

"I believe that the commission will find our policies, procedures and the manner in which these individual cases have been handled to be fully satisfactory," says Tom Closson, chief executive of the UHN. "If that is not the case, then I want to know what the commission has to say and have absolutely no reservation in saying that we will adopt all recommendations." In an internal memo sent to UHN staff last week, Closson also said he has arranged for an external review of the hospital's policies and procedures.

Schwartz has now found a new academic adviser at the University of Toronto, but she remains in dispute over access to her data. David Naylor, the university's dean of medicine and vice-provost for relations with healthcare institutions, says that Schwartz has been given much of her data and that discussions are ongoing over access to the rest. ■

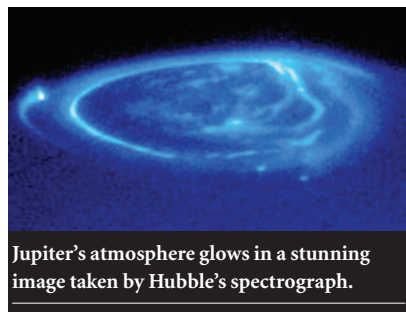
Future looks bleak for powerless Hubble device

Tony Reichhardt, Washington

One of the Hubble Space Telescope's four scientific instruments has shut down and seems unlikely to be resurrected.

Astronomers are not optimistic about repairing the Space Telescope Imaging Spectrograph, one of the telescope's workhorses. "There's some hope, but it is small," says Bruce Margon, a science director at the Space Telescope Science Institute in Baltimore, Maryland.

The spectrograph, which records specific frequencies of visible and ultraviolet light, failed on 3 August



Jupiter's atmosphere glows in a stunning image taken by Hubble's spectrograph.

owing to a faulty power unit. Project officials are currently mulling over the few options open to them. The instrument has a back-up power supply, but that has not been used since it was repaired in March 2002 and engineers are not sure whether it will work properly.

The instrument has already lasted two years longer than it was designed to do. Astronomers had come to rely on it: at present, it features in some 30% of planned Hubble science observations.

If the device cannot be revived, the onboard Advanced Camera for Surveys could provide limited spectra images. Hubble's data could also be supplemented with spectra taken from the ground.

The Cosmic Origins Spectrograph, due to be installed on Hubble during a repair mission in 2006, could perform a few of the imaging spectrograph's functions. But NASA is currently reviewing plans for the 2006 repairs and deciding whether astronauts or robots should do the job, or whether it should be cancelled altogether. The agency said this week that it would start developing plans for a robotic mission, but would not make its final decision until next year.

NASA has also awarded a study contract to a team at the Johns Hopkins University in Baltimore to see if the Cosmic Origins Spectrograph and another camera planned for Hubble could fly on a separate satellite instead. ■

Fears grow as blood stocks pass on prions undetected

Michael Hopkin, London

New infections of variant Creutzfeldt–Jakob disease (vCJD) could continue unchecked unless a sensitive screen for blood donors can be devised, researchers have warned.

Most of the 142 vCJD deaths in Britain are believed to have been caused by people eating beef products infected with bovine spongiform encephalopathy. But two people are now known to have been infected through blood transfusions containing disease-causing proteins known as prions. Researchers are concerned that in the latest case, revealed last week, the recipient of the transfusion carried the prions for years without developing the disease.

If unwitting carriers are giving blood, the spread of the disease could be perpetuated even though almost all infected beef has now been removed from the food chain. "That is a worst-case scenario, but we cannot rule it out," says James Ironside of the National Creutzfeldt–Jakob Disease Surveillance Unit in Edinburgh.

Ironside and his colleagues discovered rogue prions during an autopsy of a seemingly disease-free subject who had received a blood transfusion in 1999. They published their results on 6 August (see A. H. Peden, M. W. Head, D. L. Ritchie and J. W. Ironside *Lancet* **364**, 527–529; 2004). The elderly patient, who died of an aneurysm, had small amounts of the protein in the spleen and lymph nodes. The patient is the second person in Britain known to have been infected with disease-causing prions through a blood transfusion, but the subject in the first case went on to develop vCJD.



In the blood: donors of certain genetic types may incubate vCJD and pass it on while seeming healthy.

Researchers doubt whether current precautions, such as stripping donated blood of the white blood cells in which prions are thought to amass, are sufficient to prevent prions from being passed on by transfusions. The UK government also prohibits people who have had a blood transfusion from giving blood, but this would not prevent prions from being passed on by people infected by beef products.

All previous vCJD cases have occurred in people of a particular genetic type, but Ironside's patient belonged to a different group. Adriano Aguzzi, a neuropathologist at the University Hospital of Zurich in Switzerland, speculates that some genetic groups may act as carriers of the infectious protein without developing symptoms themselves. It will be important to see whether the rate of these infections is sufficient to maintain the epidemic, he adds.

Many more patients could contract the disease before we are able to estimate incubation periods, warns Moira Bruce, who studies prion diseases at the Institute for Animal Health in Edinburgh. Ultimately, a sensitive screen for prions in blood samples is needed, she says: "One of the holy grails is an antibody that can distinguish between normal and disease-causing prions." Bruce believes work in this area could bear fruit in the next few years.

In the meantime, says Ironside, we should be aware of the trade-off between the risks and the benefits of blood donations. "Blood is given for good reasons," he points out. "In many cases, if people do not get blood transfusions they die." ■

Security scare puts Pluto launch at risk

Geoff Brumfiel, Washington

A NASA mission to Pluto could arrive five years late and cost tens of millions of dollars more because of a security clamp-down at Los Alamos National Laboratory in New Mexico.

The nuclear-powered New Horizons spacecraft, scheduled for launch in 2006, should be fuelled by plutonium from a reprocessing facility at Los Alamos. But work at the lab shut down on 16 July, pending a security review into the disappearance of two classified computer-storage devices (see *Nature* **430**, 387; 2004). If the reprocessing facilities are not restarted soon, the mission could miss its launch window. "We're taking this very seriously," says Alan Stern, principal investigator of the project and director of the space-studies department at the Southwest Research Institute in Boulder, Colorado.

New Horizons is a \$720-million spacecraft designed to look at the Solar System's farthest planet and the Kuiper belt: a ring of icy objects beyond Neptune's orbit. Far from the Sun, the craft will depend on nuclear power to run its scientific instruments.

Under an agreement with NASA, Los Alamos is to produce 36 plutonium fuel elements. Together with 36 spare elements from



The New Horizons probe is rapidly running out of time.

the Cassini mission to Saturn, these will power the spacecraft on its nine-year journey to Pluto, says Orlando Figueroa, deputy associate administrator for programmes at NASA's Science Mission Directorate in Washington DC. To keep the satellite on schedule for the January 2006 launch date, the lab must deliver all the elements by the end of the year.

But on 4 August, Los Alamos director Peter Nanos said that it would take roughly two months for work at the lab to resume. Mission planners are racing to figure out how to work the delay into their schedule.

The two-week launch window in January 2006 is important because it would allow the craft to bounce off Jupiter's gravitational field on its way to Pluto. New Horizons could potentially launch in February 2007, but it would not be able to use Jupiter as a springboard and so would face an additional three to four years of travel time. Contingency plans drawn up by NASA put the extra costs, such as the need to run a mission control centre for longer than expected, at \$67 million.

In the coming months, mission planners will consider the implications of redesigning the mission so that it uses fewer fuel pellets. "People are looking at

lots and lots of remedies," says Stern. Adding batteries that can be used to power up instruments when the craft arrives at Pluto is one possibility. Another is that some of the spacecraft's eight science instruments could be removed, says Figueroa.

The fuel shortage is the most pressing problem currently facing the mission, but it is by no means the only challenge. Figueroa says that the launch rocket still needs final approval. And because the mission uses nuclear power, the launch-approval process is more lengthy than normal. "This is a lesson in risk management," he says. ■

Doped athletes flex muscles against drug company

Karoline Schwarzberg, Munich

Supporters of athletes who received steroids under East Germany's state-controlled doping programme are turning up the heat on the pharmaceutical company alleged to have supplied the drugs.

Pressure on Jenapharm, based in Jena, Thuringen, has been mounting since July 2003, when a former employee alleged on television that the company supplied sports scientists in the German Democratic Republic (GDR) with substances used in the doping programme. Rainer Hartwich, who now works as a physician and gives medical advice to victims of doping, also admitted to having been involved in confidential clinical doping research.

Jenapharm's management disputes the allegations, saying the company produced the drugs for medical purposes and was not involved in doping studies — a claim that will now be examined by a prominent German molecular biologist.

The extent of doping is unknown, but thousands of athletes are thought to have been given muscle-boosting steroids during the 1970s and 1980s. Some, such as Olympic

long-jumper and sprinter Heike Drechsler, say they were told these were vitamins. Steroids have many side effects, including an increased risk of cancer and infertility.



Leap of faith: gold-medallist Heike Drechsler was given 'vitamins' that were actually steroids.

Help for Victims of Doping, a Weinheim-based organization that supports athletes from the former GDR, is urging Jenapharm to pay several million euros in compensation to the group of victims. It also accuses Jenapharm, which was taken over by Berlin-based drug company Schering in 2001, of refusing to open its archives so the extent of its involvement can be established. The firm did not respond to requests for an interview.

Last month, the society enlisted the help of Werner Franke, a prominent molecular biologist from the University of Heidelberg. In the early 1990s, Franke was on a commission of the Wissenschaftsrat, Germany's influential science council, that investigated people responsible for doping. He is married to Brigitte Berendonk, an athlete who left East Germany in 1958.

During previous investigations, Franke and Berendonk unearthed documents that he is using to compile a report on Jenapharm's involvement. The report, due at the end of August, could make uncomfortable reading for the company, as Franke claims he will be able to provide evidence of Jenapharm's links to the doping programme. ■

Slipshod approvals taint Japanese animal studies

David Cyranoski, Tokyo

Animal-welfare campaigners have revealed irregularities in the way that animal experiments are approved in Japan.

The Tokyo-based Anti-Vivisection Action Network says one researcher was able to approve his own application for a project. Another application stated that 63 animals would be used, but failed to state what kind of animals were involved. "It's hard to tell whether anyone is really looking at these," says Fusako Nogami, director of the network.

The findings, taken from an initial analysis of research proposals for more than 1,000 experiments approved by 19 publicly funded research organizations, were presented to the environment ministry on 4 August. The ministry, which is holding monthly meetings on animal welfare in the run-up to an expected revision of animal-protection laws next year, says the cases could be procedural errors and do not mean that animals are being maltreated.

But Nogami says that laxity in the system could obscure abuses in labs. She wants the government to establish rules like those used in Europe and North America, where laboratories that do animal research are required to register with a government-approved body. She is also pressing for enforceable laws and an external watchdog. The section of the current animal-protection law that deals with research includes neither a penalty for offenders nor a mechanism for enforcement.

Researchers who work with animals deny Nogami's claims about possible abuses, insisting that they stick to high welfare standards that are well regulated by their institutions. But the Science Council of Japan, which represents some 1,500 scientific and academic societies, recognizes that legislation needs to be tightened up. In a report issued on 15 July, the council recommended the creation of national guidelines and an external evaluation system to ensure that they are implemented.

The environment ministry is sceptical about the proposal. A spokesperson pointed out that the research section of the animal-protection law can only be changed with the agreement of the education, health and environment ministries — a tricky feat for Japan's bureaucratic government. "It would be very difficult," says the spokesperson. ■



Free for all: the Antarctic Treaty cannot protect the frozen continent from bioprospectors.

Stiffer rules required to stop commerce milking Antarctica

Emma Marris

The Antarctic could be damaged by unregulated commercial research unless the laws covering the area are updated, polar scientists have warned.

The South Pole is governed by the Antarctic Treaty on principles of peace, openness and international cooperation — more free love than free enterprise. The agreement says nothing about how the area's flora and fauna can be exploited, prompting delegates at last month's meeting of the Scientific Committee on Antarctic Research to call on politicians to rethink the treaty.

Researchers who attended the conference, held in Bremen, Germany, from 25 to 31 July, said that commercial companies are taking an increasing interest in Antarctica. Organisms at the pole must employ complex strategies to survive, such as using antifreeze molecules to keep their insides fluid in sub-zero temperatures. Over the past five years, several patents have been granted on these and other potentially useful substances.

But such patents are not governed by the treaty, which states only that any scientific findings should be "freely available". Some fear that companies may patent an organism and prevent other scientists from working with it. "That would turn around and bite basic research," says John Priscu, a microbiologist from Montana State University in Bozeman, who attended the Bremen meeting. "It's a last frontier down there and we don't want to lock up its secrets."

The Antarctic scientific committee, which provides advice to the 45 nations that have ratified the Antarctic Treaty, cannot rewrite the rules. But Priscu says he plans to press the matter when the treaty nations

meet next summer, and he has the support of David Walton, head of the environment and information division at the British Antarctic Survey in Cambridge, UK. Before then, Priscu and Walton will draft a report on the potential scientific consequences of intense bioprospecting.

Besides the problem of patents, they worry that harvesting samples could do ecological damage. Antarctic organisms are particularly vulnerable, because they grow so slowly. Priscu studies the mats of microbes that grow on the floors of iced over lakes. These are the old-growth forests of Antarctica — it may take them up to a decade to double their size — yet researchers haul them off by the bagful for sampling. "I am concerned that these fragile mats are being destroyed," says Priscu. He plans to recommend a census of all biological material collected, by weight and area, which may encourage limits to be set.

This is not the first time researchers have warned of the dangers of bioprospecting in Antarctica. At last year's meeting of Antarctic Treaty states, the committee saw a report from the United Nations University in Tokyo that warned of a "free-for-all" and a "gold-rush" mentality if rules were not put in place. But no action was taken.

The most obvious answer, say many polar researchers, would be to adapt the guidelines of the UN Convention on Biological Diversity, which is designed to ensure that countries undertake bioprospecting in a sustainable fashion. But Walton points out that following such ready-made rules may not be possible; the United States has not ratified the convention and may refuse to abide by it in the Antarctic. ■

Doctors seek lost data on Alzheimer's vaccine

Alison Abbott, Munich

Data on the first experimental vaccine for Alzheimer's disease are being lost because the patients who were immunized are no longer being monitored.

Trials of the vaccine were abandoned for safety reasons in January 2002, but researchers in Switzerland say that much could still be learned from the patients who received the shots. "A unique biological resource" is being wasted, says Roger Nitsch, a neuroscientist from the University of Zurich who worked on the study.

The trial was launched in 2001 by the San Francisco arm of drug company Elan, which recruited 372 Alzheimer's sufferers in the United States and Europe. Participants received between one and three of six planned inoculations before evidence of brain inflammation halted the study.

The premature end of the trial dealt a blow to the concept of a vaccine for Alzheimer's, but the approach remains one of the few potentially viable options for controlling the disease (see *Nature* **422**, 370–372; 2003).

Although Elan followed the subjects until December 2002, once the trial had stopped neither the company nor any of the participating research centres were obliged to follow the patients. Of the 28 groups who provided patients for the trial, only Nitsch and his colleagues continued to study the subjects. His group has intriguing data on his patients, who now number 13, and says it would be a "great opportunity" to study other subjects from the trial.

The lack of follow-up is not unusual for this level of clinical trial. But some researchers say a longer assessment should have been made in this case, because of the long-term effects of

immunization. At present, no agency has the power to intervene to ensure such studies happen, although the US Food and Drug Administration can insist on the creation of long-term follow-up registers for some of the treatments it licenses, such as gene therapies.

Elan did not respond to requests for an interview, but the company last month presented results of its limited follow-up at the International Conference on Alzheimer's Disease and Related Disorders, held in Philadelphia. About a fifth of those immunized developed antibodies to amyloid- β protein, the main constituent of the destructive brain plaques that occur in Alzheimer's. But an alarming and unexplained brain shrinkage, averaging 6% in volume, was also seen in these patients. In addition, they showed no clear-cut improvement in cognitive function.

The results from Nitsch's group, also discussed at the meeting, were more positive. The researchers said that patients regained some of the lost brain volume in the second year after vaccination and, using a more sophisticated test than that used by Elan, they said that they had found a correlation between cognitive improvement and the level of antibodies to amyloid- β .

Nitsch is now using this technique to look for a similar link in blood samples taken by Elan from all trial patients before the study was halted. He also plans to contact all the participating centres to ask whether they will collaborate on similar tests. But each new study will require fresh ethical approval and the sourcing of new funds — exercises that could be difficult for individual researchers such as Nitsch to coordinate. ■



Long-term assessment of an Alzheimer's vaccine suggests it may improve performance in cognitive tests.

BSP MENDEL/SPL

Petition fails to reverse massive price rise for AIDS drug

Erika Check, Washington

The US National Institutes of Health (NIH) has declined to impose a price cut on a key AIDS medicine.

The refusal to accede to complaints about drug costs, made after drug firm Abbott Laboratories hiked the price of one of its AIDS medications by more than 400%, won praise from research universities.

"Pharmaceutical pricing is an issue, but marching in on a company's patent is not the appropriate way to deal with it," says Patrick White, director of federal relations at the Association of American Universities in Washington. The association feared that NIH intervention would deter companies from commercializing scientific findings.

The NIH's decision stemmed from a

petition filed in January by Essential Inventions, a Washington-based patent watchdog (see *Nature* **429**, 6; 2004). The group was protesting about the cost of Abbott's Norvir, which is used to boost the effectiveness of other AIDS therapies. Abbott, based in Abbott Park, Illinois, raised the price in December from \$1.71 to \$8.57 a day. Critics alleged that the move was designed to push patients towards Abbott's combination pill, Kaletra, which contains Norvir but was not subject to the price increase.

Essential Inventions asked the NIH to invoke laws to make the drug available in generic form. The agency can do this to patented inventions developed with NIH funds if they are not being made available under "reasonable" terms.

But on 4 August, Bonny Harbinger of the NIH's Office of Technology Transfer said that Norvir is already available to the public on reasonable terms, as it is marketed by Abbott and prescribed by physicians. "We think the issue of drug pricing is appropriately left for Congress and the administration to address legislatively," Harbinger said.

The ruling has angered AIDS patients, watchdog groups and some politicians, who vowed to fight for lower drug prices. They add that the Norvir debate has put the issue of the escalating costs of prescription drugs on the political agenda in the run-up to this November's presidential election. "This will encourage patent owners to price their products more aggressively," says James Love, president of Essential Inventions. ■

Indian company seeks US approval for generic AIDS drugs

Washington The first attempt to get generic AIDS drugs approved by the US Food and Drug Administration (FDA) was announced on 2 August.

The Indian pharmaceutical company Ranbaxy is seeking approval for its generic antiretroviral drugs, which cost less than brand-name versions, under an FDA scheme announced in May. The process seeks to approve generic AIDS drugs for distribution through US AIDS-relief programmes in developing countries.

The policy announcement followed criticism of the United States for refusing to distribute generic drugs through its AIDS programmes. Ranbaxy says it hopes to submit its products for approval by the end of the year, and the FDA plans to process applications within six weeks.

Three of Ranbaxy's drugs were removed from the World Health Organization's list of approved drugs last week owing to problems with facilities that had been used to test them. Ranbaxy says that the problems should not affect its application to the FDA, as these facilities will not be used to make the US drugs.

Women go to bed to give weightlessness a chance

London John Lennon and Yoko Ono only managed one week in a hotel in Amsterdam. But 24 women will stay in bed for 60 days next year, not to give peace a chance, but as part of an experiment to simulate extended periods of weightlessness on female astronauts.

The volunteers will not be allowed to stand up, even for meals or to wash. They must remain supine, tilted head-down six degrees below horizontal, for two months.

The experiment, which will take place in Toulouse, France, in early 2005, is part of a study by the European Space Agency, NASA and the French and Canadian space agencies to explore the role of nutrition and physical exercise in countering the adverse effects of long-duration weightlessness on female astronauts. Relatively little is known about how females are affected by weightless conditions because few women have flown in space and most previous ground-based studies have been carried out on male volunteers.

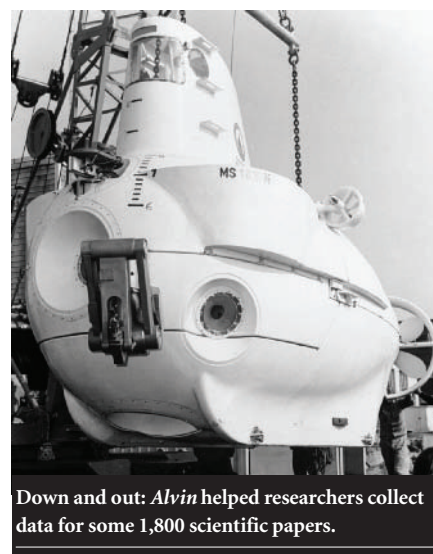
Belarus deports British Chernobyl researcher

London A British researcher studying the after-effects of the Chernobyl nuclear disaster has been expelled from Belarus and banned from returning.

Alan Flowers of Kingston University, London, has been visiting the former Soviet republic regularly for about ten years to study the impact of the 1986 nuclear-reactor meltdown on agriculture. Areas of Belarus close to the reactor, which is in neighbouring Ukraine, remain evacuated because they still suffer from high levels of radioactivity.

Flowers says he was intercepted by a KGB official in Minsk, the Belarus capital, on 29 July and told that his visa was being revoked. He is now banned from entering Belarus and Russia for five years.

The authorities have not given a reason for Flowers' expulsion, but there has been press speculation that the Belarus government was wary of the implications of his research. Some journalists have alleged that Russia seeded clouds that passed over the reactor so they would shed rain on Belarus and keep contaminated material away from Russia. But Flowers thinks it is more likely that the government objects to his work with youth groups that promote democracy and debate.



Down and out: *Alvin* helped researchers collect data for some 1,800 scientific papers.

Retirement of *Alvin* provokes deep feelings

Washington After 40 years of service, the deep-sea exploration vehicle affectionately known as *Alvin* is being retired. *Alvin*'s replacement, which has yet to be named, is due to be completed in 2008, and will be capable of diving to a depth of 6,500 metres, giving it access to 99% of the sea floor.

Alvin, which could dive to 4,500 metres, had its big moment in the late 1970s, when researchers surveying Pacific Ocean ridges found a whole ecosystem that was based not on solar energy, but on chemical energy from volcanic vents.

As well as being able to go deeper, the new submersible will have other advantages over *Alvin*. For example, a variable ballast system will allow it to work at multiple depths during a single dive. "This will allow it to do more mid-water work," explains Robert Detrick, of the Woods Hole Oceanographic Institution in Massachusetts, where the \$23-million sub is to be based.

Korean science set for government promotion

Tokyo Science in South Korea is to become a higher political priority with the impending elevation of the role of science minister to the post of deputy prime minister. The new position, which could be created as early as next month, is expected to go to the current science minister, Myung Oh.

The new government line-up will feature three deputy prime ministers, responsible for science, education and industry, respectively. Korean researchers hope the move will bring in more cash for research.

"There will probably be some long-term projects with major scientific infrastructure," says an official in the budget and planning ministry. Nanoscience and the interface between information technology and biotechnology are likely to be priorities.

Rising shuttle costs put NASA projects at risk

Washington Getting the space shuttle off the ground again will cost twice as much as NASA predicted just a few months ago, the agency's latest report says. In January, NASA estimated that 15 fixes recommended by the Columbia Accident Investigation Board, plus additional safety measures, would cost \$600 million for the three years until 2005. But an implementation plan for returning the three remaining shuttles to flight nearly doubles that estimate to about \$1.15 billion.

The increase adds to worries that NASA has failed to fully grasp the scale of the work needed (right), and will have to scrap or delay other projects unless Congress provides more money. Whether such funds are forthcoming will be



decided after the presidential election. The fixes required range from ensuring that foam insulation does not fall off the external fuel tank, as it did on Columbia's last flight, to finding ways to repair a damaged wing in space.

Noah's flood

GETTY IMAGES

Did a great flood once surge into the Black Sea, forming the basis of a Biblical tale? Quirin Schiermeier investigates a computer model that has added weight to the idea.

It sounds like a question more suited to the history of religion than science. Yet it is the driving force behind a whole field of geological research. Could a real event have inspired the Judeo-Christian story of Noah's flood — a deluge lasting 40 days and 40 nights that drowned our sinful predecessors, save one couple and their family, who took refuge with their zoo aboard a wooden ark?

The possibility has its roots in a paper published seven years ago by marine geologists William Ryan and Walter Pitman of Lamont-Doherty Earth Observatory in Palisades, New York¹. The event that became Noah's flood in the Bible, they hypothesized, was actually a massive flood of the Black Sea basin, which until roughly 8,000 years ago held a large freshwater lake. Then, cataclysmically, sea water burst through a natural dam blocking the narrow Bosphorus Strait and raised the level of the lake some 100 metres in just a few years, inundating Neolithic settlements along its shores. Ryan and Pitman subsequently popularized the idea in a book².

This 'sudden infill' hypothesis has since

become the subject of a prolonged argument between geologists, palaeontologists, oceanographers and archaeologists. Although most of them are not interested in the Biblical part of the question, they would very much like to know whether such a dramatic event happened. And despite an enormous amount of fieldwork, the jury is still out.

Now a group of geologists has taken a novel approach. Previous efforts searched for archaeological evidence of disrupted settlements or for geological evidence of the encroachment of the sea. The new approach, conceived by Mark Siddall, a young oceanographer until recently at Southampton Oceanography Centre, UK, and now at the University of Bern in Switzerland, begins in the lab with computer models of how a massive flood would transform the Black Sea basin. These results are then compared to existing geological features. And lo: most of the predicted effects seem to exist³.

Until now, most work on the flood has concentrated on its timing, rather than on its actual dynamics. But as Siddall is not concerned about whether the flood actually

influenced writers of the Bible, he says that it doesn't really matter when it occurred. More interesting, he says, is what such an extreme event would have looked like to an observer standing on the edge of the Bosphorus, and what permanent records it would have left on the sediment.

Siddall developed a fascination with the question as a PhD candidate in Southampton. Two years ago, he flew to New York to visit Ryan with some preliminary data in his luggage. Encouraged by Ryan's enthusiasm, he expanded his work.

Trained as a modeller, Siddall plugged into his computer an idealized model of the Black Sea area as it was thought to have been some 10,000 years ago, at the end of the last ice age. Back then, only a small strip of land — the now submerged Bosphorus sill — separated the Black Sea basin from the Sea of Marmara, which opens into the Aegean and the Mediterranean beyond.

At that time, it is generally agreed that the water level of the Black Sea lake was probably



Results from Mark Siddall's computer model of the Black Sea match observed geological phenomena.

D. RÄTZ

50 to 150 metres lower than the sea outside the sill. Then, fed by melting glaciers during the warming trend of the Holocene period, the Sea of Marmara began to rise. About 8,400 years ago, it reached the height of the sill and began to spill over^{4,5}

But at this point things become more blurry. Did the sill break catastrophically, allowing trillions of litres of water to flood through in a massive deluge? Or was it a more gradual event that slowly carved out the Bosphorus Strait — a channel three kilometres wide that now connects the two bodies and marks the border in Turkey between Europe and Asia?

To find out, Siddall ran a series of trials through his computer model in which he allowed the water to flow at various speeds and watched the effects on geology in the shelf, which is the portion of the basin that was covered by the flood. This sounds straightforward, but the model involved extremely complex mathematics. Simulating the ‘hydraulic jump’ — or rapid inflow of water — after a dam break required equations usually used in the lab to study fluid flow over idealized surfaces. They had never before been adapted to such intricate landforms.

Plotting a course

If water entered the basin gradually, the Coriolis force produced by Earth’s rotation would have deflected the northward flow to the right, Siddall says. A rapid flood, on the other hand — triggered by an earthquake or by rapid erosion of the Bosphorus sill — would have had a different effect. Siddall’s model shows that the powerful jet of water released by such a dam break would defy the Coriolis force by its very momentum, and would be free to take a random course.

As it happens, this latter suggestion matches up with a mysterious sharp left turn of a submerged channel that geologists have discovered in the Black Sea shelf. Seeing the

channel for the first time on a map was something of a eureka moment for Siddall. “Ryan showed me a map the Turkish Navy had made of the channel, and I said, ‘Wow — this is absolutely consistent with what the model tells us,’” he recalls.

The fast-moving current could also have formed some of the striking seafloor features seen at the mouth of the strait and in the open Black Sea, such as a number of hills and sand waves some 2,000 metres down. These sandy hills, several kilometres long and up to several hundred metres high, are likely to be the result of ‘turbidity currents’ — strong undersea flows of sediment — triggered by the moving jet.

The model also shows that the sudden flood could not have taken place in anything like the 40 days specified by the Biblical tale. According to Siddall’s calculations, about 60,000 cubic metres of water per second must have flowed into the Black Sea basin after the sill broke — more than 20 times the flow of the Niagara Falls. But even at that rate it would have taken 33 years to equalize water levels in the Black Sea and the Sea of Marmara. This is in contrast to Ryan’s much simpler flow models, which originally put the filling time at just three years.

This discrepancy does not bother Ryan, because a rise of 150 metres would still have been a notable event, even if it took 33 years. “If a model predicts observations successfully, we should pay attention to it,” says Ryan. “It adds some very important information about the dynamics of the Black Sea infill.”

Still, some scientists are uncomfortable with the whole question, because it seems to have originated from Ryan’s 30-year obsession with Noah’s flood and his eagerness to popularize the idea. In particular, they question the whole notion of basing a hypothesis on the

untestable proposal that a legend was inspired by a geological event.

“It’s all very far-fetched,” says Ali Aksu, a geologist at Memorial University of Newfoundland in St John’s, Canada, whose research suggests that the Black Sea is not likely to have filled so dramatically⁶. “Science should not be mandated by reading religious texts. That’s not the way to make hypotheses.”

And even some who are less bothered by the origins of the idea caution that the catastrophic flood is far from proven. Namik Çagatay, a geochemist and head of marine geology at Istanbul Technical University, points to his own analysis of sediments from the Black Sea’s coastal plain, which suggests that the ancient lake was only 18 metres lower than the present sea level⁷.

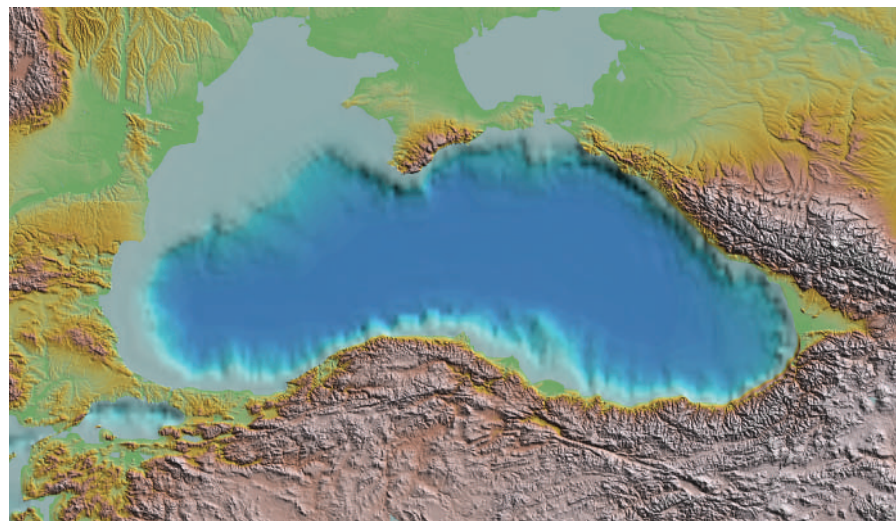
Sea change

But most researchers concede that the debate might not be so vibrant today had it not been for the flood of interest and cash brought to the field by Ryan and Pitman’s book and the subsequent media coverage. “I do believe that controversies such as this may help you get better money from funding agencies,” says Aksu. And Siddall’s work, which Aksu calls “solid physical oceanography”, shows how new interest in the field can bring a fresh perspective.

Siddall agrees that there is enough real science related to the flood hypothesis to justify talking about links to Bible stories. “Oceanographers must have a natural interest in extreme events, if only to test the robustness of their tools and methods,” he says. “If we can’t resolve the occurrence of such a huge flood, then what can we resolve?”

— Mark Siddall

“Oceanographers must have a natural interest in extreme events — if we can’t resolve the occurrence of such a huge flood, then what can we resolve?”



Now part of the Black Sea, the areas shown in grey were dry land until a flood some 8,000 years ago.

If nothing else, the new models have pointed Black Sea researchers to the next step in their fieldwork, says Liviu Giosan, a marine geologist at the Woods Hole Oceanographic Institution in Massachusetts and one of Siddall’s collaborators. For instance, they should look again for discontinuities in sediments along the shelf, which might provide evidence for a rapid rise in sea levels, he says. Sediment cores now being examined by several groups may also provide evidence for or against a cataclysmic event. Although none of this can directly link the Black Sea to the Bible story, clearly the book on Noah’s flood is not yet closed. ■

Quirin Schiermeier is *Nature’s* German correspondent.

1. Ryan, W. B. F. *et al.* *Mar. Geol.* **138**, 119–126 (1997).
2. Ryan, W. & Pitman, W. *Noah’s Flood: The New Scientific Discoveries About the Event That Changed History* (Simon & Schuster, New York, 2000).
3. Siddall, M., Pratt, L. J., Helfrich, K. R. & Giosan, L. *Paleoceanography* **19**, PA1024 (2004).
4. Major, C., Ryan, W., Lericolais, G. & Hajdas, I. *Mar. Geol.* **190**, 19–34 (2002).
5. Major, C. O. *et al.* *Geophys. Res. Abstr.* **6**, 06050 (2004).
6. Aksu, A. E., Hiscott, R. N., Yasar, D., Isler, F. I. & Marsh, S. *Mar. Geol.* **190**, 61–94 (2002).
7. Görtür, N. *et al.* *Mar. Geol.* **176**, 65–73 (2001).



Fine words: aspiring writers are lectured by film luminaries (clockwise from top left) Frank Spotnitz, Alex Singer, Martha Coolidge and Christopher Vogler.

Hollywood or bust

Last month, a handful of scientists who have toyed with the idea of writing for the movies were given a masterclass by Tinseltown's finest. Jonathan Knight joined them.

In the 1997 disaster movie *Volcano*, whole sections of Los Angeles are demolished by lava. For advice on the film, the production crew turned to Christopher Vogler, one of Hollywood's top story consultants. "Lava actually makes a tinkling sound like glass as it cools," Vogler says, "but they wanted it to roar like a freight train. Any volcanologist who saw the movie probably thought it was a comedy."

On a bright summer morning in mid-July, Vogler recounted this anecdote to a group of 15 scientists. They had come to Hollywood from all over the United States and from various scientific disciplines for a weekend workshop on screenwriting. Their goal was to learn how they could help improve the image of science and scientists in the movies. "You face an uphill battle with this stuff," Vogler continued. "But it's a good fight."

The scientists sat around a large conference table in the sunny library of the American Film Institute (AFI), a writing and directing school nestled on a hillside just

minutes from the world's largest film studios. Shelves of scripts lined the walls and dozens of film stars — including Michael Caine, Sean Connery and Audrey Hepburn — gazed down from posters as Vogler and a cast of Hollywood luminaries gave their best advice on everything from writing a good script to getting a production company to read it. At times the challenges seemed demoralizing, and yet by the end of the workshop, the scientists were, if anything, inspired.

"It gave me the motivation to consider screenwriting as a serious hobby, if not a potential new career," commented Leo Cheng. Currently, Cheng works at the Jet Propulsion Laboratory (JPL) in nearby Pasadena as a science-planning engineer for Cassini, the spacecraft now orbiting Saturn. "I learned a great deal," he said.

The portrayal of science and scientists in the mass media are two separate but related

issues, both of which have long been a source of irritation for real-life researchers. We all know that scientific accuracy generally falls by the wayside when it might interfere with dramatic effect. How exciting is a silent explosion in outer space, after all? And if it took decades instead of days for global warming to flood New York, the current blockbuster *The Day After Tomorrow* might have made less of a splash. Still, this leaves many scientists worried that such movies serve only to misinform a public whose

knowledge of science comes mainly from the big and small screens.

But it is the issue of how scientists themselves are portrayed that brought many of the participants to the workshop. Diandra Leslie-Pelecky, a materials scientist at the University of Nebraska in Lincoln, fears that media stereotypes are turning generations of children off any thought of a career in research. In

"In film and television, scientists are often quirky, nerdy, obsessed, reclusive, self-important and not infrequently mad. These are not character traits that appeal to kids."



Waiving the science: *The Day After Tomorrow* favours drama over accuracy for the flood of New York.

film and television, scientists are often quirky, nerdy, obsessed, reclusive, self-important and not infrequently mad. These are not character traits that appeal to kids, she says.

Leslie-Pelecky has had numerous conversations with children about their views on scientists. She is leading a project funded by the National Science Foundation that is investigating the impact of the media on children's attitudes. She has encountered many children who believe that the researchers who have visited their elementary schools aren't the real McCoy. "They might say the person was too 'normal' or too good-looking to be a scientist," Leslie-Pelecky told *Nature*. "The most heart-breaking thing is when they say, 'I didn't think he was real because he seemed to care about us.'"

Much misunderstood

For others at the workshop, improving the public's perception of scientists was of personal importance. "Whenever I meet a non-scientist, I'm afraid they misunderstand what I am," says Michael Strong, a graduate student in biochemistry at the University of California, Los Angeles.

About five years ago, such concerns led William Wulf, president of the National Academy of Engineering, to call his friend Alex Singer, a television director. "He was worried about the decline in the number of young people entering the sciences, and he wondered whether the media could be

having an effect," recalls Singer, whose credits include *Hill Street Blues*, *Star Trek: The Next Generation* and *MacGyver*.

The discussion rumbled along for some time until the idea for the workshop finally gelled. Martin Gundersen, a physicist at the University of Southern California in Los Angeles and an acquaintance of Singer's who has served as a science adviser on Hollywood films, helped to arrange funding from the Air Force Office of Scientific Research. The show was on.

Lights, camera, action

More than 50 scientists applied to participate when the workshop was announced in May. The 15 chosen had all demonstrated an interest in screenwriting, Gundersen says, and each came to the workshop armed with a script idea.

Carl Carlsson, an environmental engineer with a Houston power company, described his idea for a screenplay about a 14-year-old girl growing up on the Moon. James Brown, head of discovery bioinformatics at GlaxoSmithKline's facility in Upper Providence, Pennsylvania, brought along his bioterror-themed idea in which recent advances in gene therapy are twisted by terrorists to do more sinister things. And Ron Garret, who designs computer modules for martian rovers at JPL, came with a draft script called *The Cure*, in which a mysterious disease changes the lives of three

molecular-biology graduate students.

"I was very intrigued by all of your ideas," said George Walczak, a screenwriter and graduate of the AFI. But as the scientists thrashed out their ideas with the instructors, many of them learned that they were missing key elements — without which their screenplays had little chance of success.

The formal structure of a film script is fairly standard. It runs to around 120 pages, roughly a minute per page, and is generally divided into three acts of about 30, 60 and 30 pages, respectively. The script details the action and dialogue but leaves most of the specifics of shooting, such as camera movement, up to the director.

The surprise for many of the scientists was the overwhelming importance every one of the workshop's instructors placed on character-based narratives. The message, in a nutshell, was that the main theme of a dramatic film cannot be Mars exploration or evolution or even a scary virus. Instead, it has to be a protagonist and his or her 'journey'. Any scientific content is incidental.

Hero worship

So, for example, the 1997 film *Contact*, based on a book by the late astronomer and science popularizer Carl Sagan, would not have worked had it just been about the search for extraterrestrial intelligence. It succeeded as a film because it was about an astronomer — played by Jodie Foster — who dreamed her whole life of reaching the stars and who battled numerous obstacles on her journey. This eventually took her to a distant corner of the Universe to meet alien beings.

A good script is essentially the classic hero's journey, key elements of which can be found in everything from Homer's *Odyssey* to aboriginal folk tales. The stages of this journey are outlined in a number of scriptwriters' guides, including Vogler's seminal book, *The Writer's Journey: Mythic Structure for Writers*, which has become a key reference for many Hollywood screenwriters.

By this test, many of the scientists' script ideas were not yet fully formed, and the instructors had a number of ideas for improving them. Taking Brown's bioterror idea, Vogler suggested allowing the central character, a cancer epidemiologist named Paul Edwards, to express a wish early in the film. It could be to win the Nobel prize or to overcome some character flaw, Vogler said. "When the protagonist makes a wish, the story immediately wakes up and pays attention," he explained. The audience then has a goal by which to mark the character's progress.

The challenge, then, is how to bring science into a script that is driven by the story of its protagonist. "I read things in *Science* and *Nature* every week that are absolutely fascinating," said Tom Katsouleas, a plasma physicist at the University of Southern California, citing a *Nature* News Feature

published in February about the last universal common ancestor, the theoretical first cell that gave rise to all cellular life on Earth (see *Nature* 427, 674–676; 2004). He asked the instructors: “How do we put something like that into a script about character?”

Frank Spotnitz, who for four years was executive producer of the TV drama *The X-Files*, suggested that the protagonist might have to recreate this creature to save the planet. “If it were for *The X-Files*, I’d think about how it could start killing people,” he added.

Life-or-death situations are essential. The stakes must be high in a screenplay, whether for film or TV. Walczak calls this the “law of thermodynamics.” “If you aren’t sure what the stakes are, find out what your main character cares most about in the world,” he urged the scientists. “Those are your stakes.”

Lab liabilities

The realities of screenwriting are sure to disappoint those scientists who wish the world at large could get a more realistic glimpse of what their daily lives are like. As with many other professions, the daily drudgery of research does not make for good movie action.

Spotnitz told the group he toyed with the idea of showing the details of an experiment on *The X-Files*. The crew filmed a half-hour segment of Agent Scully working in the laboratory, but killed it and rewrote the script after everyone agreed it was far too boring. Singer concurred: “Most scientific work is unfilmable.”

On top of all this, the screenplay writer has to consider a factor that book writers are free to ignore: cost. For instance, making a character age 15 years over the course of the film might be important to the story, but the producer who reads it will immediately think of the hundreds of thousands of dollars it costs to artificially age the actors and shift the time period of the sets and costume. Similarly, writing “a freight train passed by” has enormous filming implications because of the time and expense required to back up a train for a second take. Producers care about cost, and writers ignore this at their peril.

Even after hearing this, the scientists in the workshop remained enthusiastic. “Show me the constraints and I can work within them,” said Leslie-Pelecky. “That’s what I do for a living.”

But there was more to come. Writing a good script is only the first challenge. In particular, a producer then has to be convinced to buy it. The usual process involves finding somebody who knows somebody at a production company just to get the script in the door. Then it goes to ‘readers,’ who are often young, aspiring screenwriters themselves.



Journey’s end: in *Contact*, Jodie Foster’s character realizes her dream of reaching the stars, while portraying the life of a radio astronomer reasonably accurately.

Their safest bet is to say ‘no,’ rather than risk getting egg on their face for passing a lousy script on to a production executive. “They reject most of what comes in, and nine times out of ten the executive rejects what the reader passes,” said Spotnitz.

Hollywood’s dream factory can afford to be so picky because of the sheer number of scripts that come in every day. So far this year, 40,000 script ideas have been registered with the Writers Guild of America, according to Singer: “And this is a slow year.”

Killer blows

Once a company has bought the option to make a movie, the producers are likely to want the writer to rewrite the script to their specifications. If they aren’t pleased with the new version, they may kill it. If they are, they still need to find a director interested in making the film. If they don’t, again the script dies.

There are so many ways for a script to die after it is optioned that many professional screenwriters don’t have a single movie credit. But they keep at it. “There are a lot of people who make very good money writing very good scripts that never get made,” Singer told the scientists.

Despite all this, there is hope for the future of science in film. For one thing, some directors are starting to care about scientific accuracy. Director Martha Coolidge, who made a presentation at the workshop on Sunday afternoon, hired Gundersen as a science adviser for her 1985 comedy *Real Genius*, about some brainy kids who develop a new high-powered laser. “The labs had a very real look, we even had a laser on the set,” Gundersen said. “Comments from people in the field were uniformly positive about the accuracy of the science.”

More recently, *Contact* gave a reasonably

accurate portrayal of the work of radio astronomers. *Apollo 13*, the 1995 film about the ill-starred Moon mission, contained numerous accurate details about space travel. Such attention to detail raises the bar for other films, Singer claimed: “You can’t make a stupid movie about the Moon any more.” Even children’s movies such as *Finding Nemo* have won plaudits for their attention to scientific detail (see *Nature* 427, 672–673; 2004).

Workshop participant Valerie Weiss, who recently completed a PhD in biochemistry and has directed and produced two short films, is optimistic about the prospects of getting more and better science into the movies. Unlike most of the participants, Weiss has already turned all her attention to film. She founded a film programme while she was a graduate student at Harvard Medical School and moved to Hollywood in 2001 to pursue film-making full time. She now has a script idea for a movie about a scientist, but asked *Nature* not to reveal it in detail. “You can be scooped in film-making just like in science,” she said.

Whether any of the ideas discussed at the workshop will end up on the big screen remains to be seen, but the next steps are in the works. Gundersen is thinking about holding a five-week workshop at the AFI next year, at the end of which each participant would have completed a script. Joe Petricca, vice-dean of the AFI Conservatory, has offered to organize screenwriting instructors or professional readers to provide feedback for any of the current participants who can submit the first 30 pages of a script by 1 October.

Brown, who is now fired up to work on his bioterror script, plans to take Petricca up on this offer. He feels better prepared knowing what challenges lie ahead. “It’s daunting trying to get a paper into *Nature*, too. Yet people do it,” he said after the workshop. “This is just the same.”

Jonathan Knight writes for *Nature* from San Francisco.

Migration won't make Chinese deserts bloom

Diverting water is not enough. Fragile land such as Xinjiang needs care, not exploitation.

Sir— In the Commentary “Agriculture of the future” (*Nature* **428**, 215–217; 2004), T. C. Tso argues that the Chinese government should use the country's three most important resources — people, land and water — to their best effect. We agree with this general principle. But we have three reservations regarding turning the Xinjiang region into the “California of China”.

First, the author gives the impression that Xinjiang is largely undeveloped and has “abundant resources”. Yet the area is home to 11 people per km², far exceeding the international population-density standard for arid areas of 7 people per km². Some 95% of the population live in oases totalling less than 60,000 km². The population density in oases, at 292 inhabitants per km², reaches that of many coastal cities. Far from being “undeveloped”, the area is densely populated and would have difficulty absorbing new migrants. Also, far from having “abundant resources”, it already experiences pressure on water sources. Large-scale agricultural expansion will

require large amounts of water that are currently not available. Tso argues for the diversion of water from southern China, but this has the potential to create great ecological harm and international conflict, as discussed by A. Chen and C. Chen (*Nature* **429**, 501; 2004).

Second, the arid Xinjiang region is more ecologically fragile than Tso suggests. Desertification has increased as a direct result of overgrazing and unsustainable farming practices. During the past 50 years the area of desert has increased by 4,210 km², at an average rate of 84.2 km² per year, while lakes have decreased from 9,700 km² to 4,748 km². Many famous lakes have actually disappeared.

The area's main river, the Tarim, has been steadily dwindling for 30 years as a result of dam building and irrigation upstream. This in turn has reduced the growth of vegetation, including ancient poplar groves, that had long formed a barrier between two deserts. It is now possible that the Taklimakan desert and the Kumtag desert will merge. Each year the

frequency and intensity of dust storms increase, and the effects are felt as far away as Beijing, Tianjin, Korea and Japan. If more people move into Xinjiang, the problem of desertification will inevitably be exacerbated.

Finally, the large-scale migration proposed by Tso could have serious cultural implications. The Xinjiang region is inhabited by the Uighur people, the fifth largest ethnic minority in China. Immigration by ethnic Han people could alter traditional ways of life followed by the Uighur.

Development in Xinjiang requires an adequate understanding of its environment and people. Turning it into the “California of China” is not a viable option.

Wenwei Ren*, Andrew A. Saret†

**Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai 200433, People's Republic of China*

†School of Urban and Regional Planning, Queen's University, Kingston, Ontario K7L3N6, Canada

Bermuda welcomes careful prospectors

Sir— Despite the claim in your News story “Bermuda gets tough over resource collecting” (*Nature* **429**, 600; 2004), Bermuda's Ministry of the Environment has not shut down any research projects relating to biodiversity access, even on a temporary basis. New laws and regulations are under development to enhance bioprospecting, not to prevent or hinder such research activities.

Contrary to your News story, the Ministry of the Environment is not displeased with Diversa's research activities in Bermuda. The ministry greatly values the ongoing collaborations with the Bermuda Biological Station for Research (BBSR), and appreciates the station's responsibility in ensuring a proactive and consultative approach to issues of environmental access. Furthermore, Diversa, BBSR and the ministry look forward to exploring ways of expanding the bioprospecting benefits realized to date by the people and the government of Bermuda.

Jack A. Ward*, Anthony Knap†, Jay M. Short‡

**Ministry of the Environment, Government Administration Building, 30 Parliament Street, Hamilton HM 12, Bermuda*

†Bermuda Biological Station for Research,

Ferry Reach, St George's GE 01, Bermuda
‡Diversa Corporation, 4955 Directors Place, San Diego, California 92121, USA

Golden rule of economics yet to strike prospectors

Sir— Your recent News Feature “Bioprospects less than golden” (*Nature* **429**, 598–600; 2004) is the latest to document the disappointments of bioprospecting. It suggests that “new rules on ownership” might give bioprospecting a boost. Our experience leads us to doubt this conjecture. The reason commercial interests are not willing to pay much for biodiversity is that it is not worth much to them. This may seem paradoxical, given that new products can be worth billions of dollars, but our conclusion emerges from several basic principles of economics.

First, the spectacular earnings available from infrequent research successes must cover the far more common failures. Second, even in bioprospecting, costs are dominated by specialized labour and equipment rather than the chemical entities from which products are crafted. Finally, and most importantly, untested natural chemical entities are not scarce, and for that reason, they should not be expensive. A pharmaceutical researcher's

willingness to pay for a natural product lead is determined by the probability that the lead will result in a useful new product multiplied by the probability that the same product would not be developed from any of the other leads available to the researcher. When there are large numbers of potential leads, successes must either be so unlikely as to make searching unattractive or so probable as to render competing leads redundant. Either way, researchers will not pay much for any particular untested product.

Your News Feature emphasizes that rules must be agreed for the allocation of genetic resources, and this is certainly important. Yet the biggest practical problem with bioprospecting may be unrealistic expectations. Our sense is that governments and advocates in developing countries have often been misled about the value of their biological wealth. Consequently, they hold out for spectacular gains that they are unlikely to achieve, and all too often forego the more modest but more readily achievable benefits such as those that your News Feature describes as arising from Phyllis Coley's programme in Panama.

David Simpson*, Roger Sedjo†

**Department of Economics, University College London, Gower Street, London WC1E 6BT, UK*

†Resources for the Future, 1616 P Street NW, Washington DC 20036, USA

Seeking a global solution

The Copenhagen Consensus neglects the need to tackle climate change.

Jeffrey D. Sachs

To address the global challenges of poverty, climate and disease, it is essential to build effective bridges between the scientific community and policy-makers. These issues are much too complex to leave to the normal give-and-take of interest-group politics. The 2004 Copenhagen Consensus failed to build these bridges in an effective manner, I believe — but others could learn from its flawed approach.

The Copenhagen Consensus was set up by the Danish Environmental Assessment Institute, directed by Bjørn Lomborg, to identify priorities for global action regarding poverty, health, hunger and the environment. With an expert panel consisting of eight prominent economists, including three Nobel laureates, and with the backing of *The Economist* magazine, the project received considerable attention. The panel reviewed 32 proposals and ranked half of them as good or bad investments. The “very good” rating for HIV/AIDS and malaria policies, and the “bad” rating for three proposals to control climate change, generated the most publicity when the results were announced in May.

Unfortunately, the project’s headline conclusions do not withstand close scrutiny, especially the conclusion that climate control is a “bad” global investment. More generally, the conclusions offer little sound guidance to policy-makers, much less a new consensus. On the positive side, the project did help to increase public awareness of important global issues, and has produced several useful background papers.

Wrong question

The panel that drew up the Copenhagen Consensus was asked to allocate an additional US\$50 billion in spending by wealthy countries, distributed over five years, to address the world’s biggest problems. This was a poor basis for decision-making and for informing the public. By choosing such a low sum — a tiny fraction of global income — the project inherently favoured specific low-cost schemes over bolder, larger projects. It is therefore no surprise that the huge and complex challenge of long-term climate change was ranked last, and that scaling up health services in poor countries was ranked lower than interventions against specific diseases, despite warnings in the background papers that such interventions require broader improvements in health services.

It is worth putting the extra \$50 billion — \$10 billion per year — into context. Annual



Waiting for help: floods in Bangladesh last month left people in Dhaka queueing for relief supplies.

income in the world is currently about \$40 trillion, of which some \$30 trillion is in the high-income (donor) countries. So the project looked at investing a measly 0.03% of annual donor-country income to address the planet’s greatest challenges — hunger, disease, environmental degradation and instability — which are life-and-death issues for a billion or more of the world’s poorest people. The United States alone now spends almost \$450 billion per year on the military, a rise of \$150 billion in the past three years.

The Copenhagen Consensus would be more convincing if it acknowledged what the rich world has already promised. At the International Conference on Financing for Development in March 2002, both rich and poor countries adopted the Monterrey Consensus, which declared: “We urge developed countries that have not done so to make concrete efforts towards the target of 0.7% of GNP as official development assistance.” This figure is currently 0.25% of donor GNP, or about \$69 billion per year. The Monterrey goal would mean spending about \$210 billion per year, an increase of some \$140 billion per year. This is 14 times the sum suggested by the Copenhagen Consensus.

Moreover, the world has already committed to fighting disease, hunger and climate change on a bold and broad scale. The Millennium Development Goals, adopted in September 2000, call for dramatic steps to cut child mortality by two-thirds and the number of people suffering from hunger by half

by the year 2015, compared with a 1990 baseline. The United Nations Framework Convention on Climate Change already commits the world to “stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system”.

With the funds already promised, which can easily be delivered by developed nations, the world would not have had to choose between addressing specific diseases or overall health systems, or between small-scale water projects and long-term climate change. We could do both. The Copenhagen Consensus could then have usefully turned its attention towards how to accomplish those tasks, rather than whether to follow through on commitments already made.

Wrong participants

Mobilizing expert analysis to inform policy-makers is a good idea. A remarkable example is the Intergovernmental Panel on Climate Change (IPCC), which since 1988 has brought together hundreds of leading climate scientists, economists, engineers and other specialists on the issue of anthropogenic climate change. The IPCC has helped the climate-change community to identify areas of consensus and disagreement, and to improve communication across scientific and other disciplines.

Another example is the Commission on Macroeconomics and Health (CMH) of the World Health Organization in 2000–01,

which I had the honour to chair. That effort brought together more than 100 scientists, policy-makers, economists and practitioners for a two-year study on health in low-income countries.

The Copenhagen Consensus fell far short of the IPCC and CMH as a deliberative process. It failed to mobilize an expert group that could credibly identify and communicate a true consensus of expert knowledge on the range of issues under consideration. The panel included distinguished economists but no natural scientists or public-health specialists. Nine of the ten authors of background papers were also economists, as were almost all of the 20 'opponents' (commentators) on the background papers.

With the exception of Robert Fogel, the panel is mostly known for its expertise outside the areas under discussion. A panel of economists brings an important set of tools to the table, but it cannot accurately assess the social costs and benefits of alternative interventions regarding climate, agriculture, disease, water and nutrition without the input of natural scientists, engineers and public-health practitioners.

The evidence made available to the panel was also extraordinarily narrow: the panel received a single background paper and two short opponent papers for each proposal, often with highly idiosyncratic results. The project's timeline itself was far too short for the panel to gain requisite expertise, lasting only a few months in total; the background papers circulated for a few weeks, and in the final discussions, the panel had 5 days to review 32 proposals.

Wrong conclusions

Many of the panel's conclusions were at odds with the evidence under consideration, and no rationale was offered to the public. Climate change is a salient case. The background paper, by William Cline of the Center for Global Development in Washington, examined the benefits and costs of several strategies to mitigate climate change. Cline's central strategy — an aggressive global carbon tax — showed a long-term benefit-to-cost ratio of 2:1, according to the simulation model that he used. The opponents called for a more gradual approach, but both endorsed a framework of early action to regulate and limit carbon emissions, with increasing constraints in future decades.

Despite this agreement between the author and the opponents, the panel concluded that a global carbon tax is a "bad" investment. The brief description of the panel's judgement notes that "the experts expressed an interest in an alternative, proposed in one of the opponent papers", in which an initially low carbon tax rises gradually in later years. The panel also "urged increased funding for research into more

affordable carbon-abatement technologies". But these proposals were not ranked as they were not examined in detail. Had there been time to examine these proposals further, the headline conclusion might have been that "the Copenhagen Consensus panel supports a carbon tax".

The proposals regarding health services, disease control and nutrition reveal similar problems. The panel rated interventions aimed at HIV/AIDS and malaria as "very good", but rated as only "fair" a broad scaling-up of basic health services in low-income countries. Similarly, the panel rated as "very good" and "good" two targeted interventions regarding malnutrition — micronutrient supplementation and new agricultural technologies — but ranked as only "fair" two broader and more expensive interventions directed at infant and child nutrition and low-birthweight babies. The bias towards smaller, cheaper, targeted projects is clear, but runs contrary to the experts' advice in the background papers.

The basic-health proposal came from Anne Mills of the London School of Hygiene and Tropical Medicine, who wrote the Copenhagen Consensus background paper on communicable diseases. Her paper stresses the interdependence of the proposals: "both malaria and HIV/AIDS control must include a substantial component of strengthening health services if they are to be successful." Both opponents also stressed the need to strengthen health systems generally if investments on specific diseases are to reach their potential, and discussed ways to organize such spending.

As with climate change, then, the background papers and opponents' comments do not really justify the final ranking. In 2000–2001, Mills co-chaired a task force for the CMH that found a very high benefit-to-cost ratio (roughly 6:1) for a broad-based scaling-up of health services in poor coun-

tries. Her study also found that some 8 million deaths each year could be averted within a decade or so. Here, Mills again finds a high benefit-to-cost ratio (now 3.9:1) for a designated package of health interventions for a group of poor countries.

The Copenhagen Consensus panel's low ranking of the broader health interventions seems to reflect the project's two main methodological flaws: the assumption that additional funding is limited to \$50 billion over five years; and the lack of experience in health issues among panel members, which perhaps made the obstacles to a broad-based scaling-up of health and nutrition services look more daunting than they are.

Lessons for the future

The core concept of the Copenhagen Consensus is a good one: to engage expert opinion to evaluate policy options on major challenges facing the planet. This approach has been used before and should be adopted again. But the project's methodological failures should not be repeated.

First, future projects should avoid a statement of the policy problem that is likely to bias the results. Second, the expert group should have sufficient time to consider the evidence thoroughly, to consult with external experts, and to consider any proposals that seem to have consensus support. Third, the expert group should cover all relevant disciplines, including several members with long-standing professional experience in each area under examination. Fourth, initial assessments by the expert group should be widely circulated for comments, corrections and feedback before a public report is issued. In this way a true 'consensus' may be reached, and the public won't be left wondering which set of experts to trust. ■

Jeffrey D. Sachs is director of the Earth Institute at Columbia University, 535 West 116th Street, New York, New York 10027, USA.



On the right track? Flooding in India highlights the plight of vulnerable people.

P. SINGH/APP/GETTY IMAGES

Unhappy pills?

The trials and tribulations of antidepressant drugs.

Medicines Out of Control: Antidepressants and the Conspiracy of Goodwill

by Charles Medawar & Anita Hardon
Aksant/Transaction: 2004. 258 pp.
£19.25/\$34.95

Let Them Eat Prozac: The Unhealthy Relationship between the Pharmaceutical Industry and Depression

by David Healy
New York University Press: 2004. 340 pp.
\$29.95

John Cornwell

In 1994, as lawyers for the pharmaceutical giant Eli Lilly were beginning a historic legal defence of the antidepressant Prozac, researchers were descending from all over the world on Miami Beach, Florida, for the twenty-fourth annual meeting of the American Society for Neuroscience. Among the 10,000 posters on show at the meeting, hundreds displayed research on the serotonin system, with scores of these devoted to fluoxetine hydrochloride, better known as Prozac. One poster showed the work of Ray Fuller of Eli Lilly, who invented Prozac. But Fuller wasn't there. He was in Louisville, Kentucky, giving evidence in the Eli Lilly Prozac trial. It involved a man called Joseph Wesbecker, accused of shooting 20 people in his workplace with an AK47 after being on Prozac for just over a month. Eight of his victims died and he committed suicide.

This was the mid-point of the Decade of the Brain. To attend the annual neurosciences meeting was to submit oneself to a tidal wave of varied research programmes, from genetics to neuronal degeneration, from memory to neurotransmitters, from pharmacology to the visual cortex, and from brain imaging to new pain therapies. Nowhere in the world was there such a large and optimistic gathering of scientists convinced that they were about the change the world.

The decision of the House and Senate of the United States to designate the 1990s as the Decade of the Brain was not just scientific optimism, however; it was the result of lobbying by the pharmaceutical industry. The carrot had been budgetary: an estimate that some \$350 billion was being lost to the US economy each year through brain-related ills, including depression, Alzheimer's and the consequences of aggressive behaviour. The promise of major social and medical amelioration was driving and directing the rapid expansion of investment and funding. These two volumes, *Medicines Out of*

Depression hurts.

Depression isn't just feeling down. It's a real illness with real causes. Depression can be triggered by stressful life events, like divorce or a death in the family. Or it can appear suddenly, for no apparent reason.

Some people think you can just will yourself out of a depression. That's not true. Many doctors believe that one thing that may cause depression is an imbalance of serotonin — a chemical in your body. If this happens, you may have trouble sleeping, feel unusually sad or irritable. Find it hard to concentrate. Lose your appetite. Lack energy. Or have trouble feeling pleasure. These are some of the symptoms that can point to depression — especially if they last for more than a couple of weeks and if, instead, everyday life feels like too much to handle.

To help fight depression, the medicine doctors now prescribe most often is Prozac.[®] Prozac isn't a "happy pill." It's not a tranquillizer. It won't turn you into a different person.

Some people do experience mild side effects, like upset stomach, headache, difficulty sleeping, dizziness, anxiety and nervousness. These tend to go away within a few weeks of starting treatment, and usually aren't serious enough to

make most people stop taking it. However, if you are concerned about a side effect, or if you develop a rash, tell your doctor right away. And don't forget to tell your doctor about any other medicines you are taking. Some people should not take Prozac, especially people on MAO inhibitors.

As you start feeling better, your doctor can suggest therapy or other means to help you work through your depression. Prozac has been carefully studied for nearly 10 years. But remember, Prozac is a prescription medicine, and it isn't right for everyone. Only your doctor

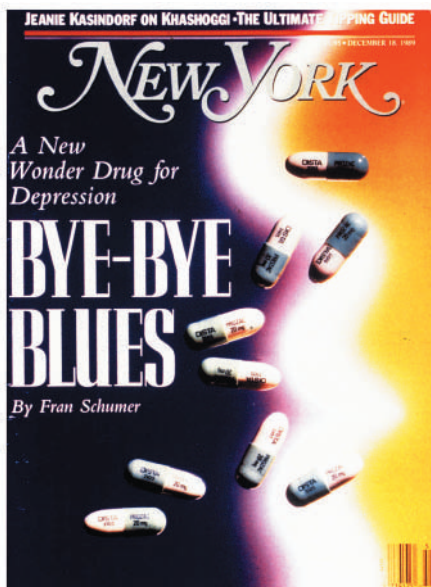
can decide if Prozac is right for you — or for someone you love. Prozac has been prescribed for more than 11 million Americans. Chances are someone you know is feeling better again because of it.

prozac
fluoxetine hydrochloride

Welcome back.

Please see important information on following page. ©1999 Eli Lilly and Company

Blue skies — or just blues? Prozac seemed like a miracle treatment for depression, but after a blaze of publicity in the media (below), things soon started to turn sour for its manufacturer, Eli Lilly.



Control and *Let Them Eat Prozac*, measure the failure of the Decade of the Brain to deliver the goods in the crucial arena of new antidepressants.

A chief benefit of new psychopharmacology was the hope that designer medications would bring an end to dependence problems. In theory, matching a chemical product to known receptor sites would avoid the unwanted side-effects of older drugs, whose discovery was mainly serendipitous. *Medicines Out of Control* tells the story of a looming crisis in the side-effect of depen-

dency on new drugs. The turning point signalled by the authors was 25 June 2003, when GlaxoSmithKline published an amendment to the prescribing instructions for the antidepressant Seroquel (Paxil). The company revised its earlier estimate of the risk of withdrawal reactions from 0.2% to 25%. In setting the scene for their thesis, the authors tell us that the prodigious increase revealed that "science was catching up with common sense".

One of the central issues of the book is the mismatch between the industry's clinical-trial systems and the anecdotal evidence of patients. The authors argue that evidence from users began to flow in earnest in the mid-1990s with the growth of the Internet. The web "had begun to change history" as users of antidepressants began to compare notes and exchange ideas on an unprecedented scale. The authors' conclusion — that the community created by new information technology should be adapted and adopted by the medical profession, especially in the case of drugs for the mind — seems not only sensible but obvious. Will the pharmaceutical industry respond? Is it in their interest to do so?

Which brings us to David Healy's *Let Them Eat Prozac*, an even more sombre story of psychopharmaceutical folk and patients. For beyond the dangers of dependence are the alleged dangers of aggression and self-harm, particularly in the case of the selective serotonin reuptake inhibitors (SSRIs), of which the most famous is Prozac. Healy,

who is a psychiatrist and prolific author of books and articles on psychopharmacology, used to prescribe Prozac happily. Like many clinicians, he was impressed: just one pill a day to bring a patient out of a wide variety of depressive complaints. Only with time did he realize that his patients were complaining of agitation and suicidal ideation. Then came that landmark trial in Louisville involving Wesbecker and Eli Lilly. The many expert witnesses revealed the alarming relationship between new brain science and the commercial goals and shortcuts of the pharmaceutical industry.

Healy's tale is complex and detailed, bringing together a huge amount of clinical-trial data and case histories. Ultimately, the book is about science, society and the power and misuse of commercial promotion. "We are facing a future of real biomedical developments," he concedes, but these are taking place in a world "in which corporate capacities to colonize the consciousness of citizens, physicians, regulators, and others outstrip their capacities to bring real benefits on stream". His investigation is impressive but is, in the final analysis, depressing, for he has no ready answers for the ills he describes. One can only hope that the benefits of commercially based psychopharmacology ultimately outstrip the catastrophes — but Healy is not prepared to concede such a conclusion. Perhaps our best hope is for vigilant and painstaking whistle-blowers, of which the authors of these two studies are formidable examples. ■

John Cornwell is director of the Science and Human Dimension Project, Jesus College, and is in the Department of History and Philosophy of Science, University of Cambridge, Cambridge CB2 3RH, UK. He is the author of *The Power to Harm: Mind, Murder and Drugs on Trial*.

The King of Siluria's journey

Murchison's Wanderings in Russia
edited by Michael Collie & John Diemer
British Geological Survey: 2004. 474 pp. £40
<http://www.bgs.ac.uk/bookshop>

Ralph O'Connor

If you had been rowing across the Volga near Kazan, western Russia, in May 1841, you might have been startled by the sound of raucous singing:

*Ah, the red sandstone! How bored am I!
I'd very well pay a thousand louis
Never again in my life to see
The new red sandstone of Tartary!*

The singer (and composer) was Roderick Murchison, the greatest practical geologist of



Geologist Roderick Murchison studied the gorge of the River Chusovaya on his travels through Russia.

his day and a key player in Queen Victoria's empire. In their search for mineral resources, Britons of the time were organizing and stamping their authority on landscapes past and present, naming periods such as the Cambrian, Silurian and Devonian after British localities and peoples. Strata had become the central concern of geology, not so much for telling the story of life — though this was a good way of selling books — as for mapping the positions of rocks and correlating them across the globe.

Murchison had become famous in the 1830s as the man who made sense of the earliest Palaeozoic strata in the Welsh borders. He now meant to give his 'Silurian system' a global reach. In 1840 and 1841, Murchison and three colleagues sped across the Russian plains, covering thousands of miles and adding new territories both to his geological empire and to the Tsar's coal reserves. Such an exhausting and exhaustive campaign naturally had its dull moments, hence the song quoted above.

His findings were written up in the hefty *Geology of Russia* (1845), and the honours he had craved since he abandoned his military career at last poured in — he was hailed as the 'King of Siluria' and as the Copernicus of the age. But intellectual ossification was quick to follow: from this point on he knew he was right, and cultivated the impenetrable formality of the retired general.

Murchison also wrote up his travels in a less formal manner. His journal, *Wanderings in Russia*, provides a continuous narrative of his 1840 and 1841 expeditions, and is here published for the first time. This handsomely produced book should be welcomed by historians and geologists alike. It is full of fascinating anecdotes about local customs, manners and rocks from St Petersburg to the

Urals. Social commentary was not Murchison's forte, but in his clipped way he provides a rich mine of specific information about the land through which he was travelling.

As a travel writer, Murchison cannot compare with his fellow geologist and Scotsman, Hugh Miller, whose marvellous Hebridean travelogue *The Cruise of the Betsey* (1845–49) has also recently been reprinted (see *Nature* 426, 19; 2003). Where Miller has the novelist's gift for character and incident, Murchison just rambles on. But once you get used to the amoebic shape of the narrative and its periodic outbursts of self-congratulation, *Wanderings in Russia* is an enjoyable read, full of humour as well as picturesque description. If is, of course, highly revealing of Murchison's own attitudes, including his affection for the Russians at a time when British 'Russophobia' was on the rise.

Michael Collie and John Diemer have done a splendid job of editing this massive text, helpfully splitting it into short sections and inserting a running commentary, with detailed maps and notes. The full-size colour reproduction of the *Geology of Russia*'s enormous geological map, tucked conveniently into a pocket at the back, deserves special praise. My only reservations concern the rather short introduction, which, despite many useful insights, is too involved in minutiae to give the general reader much of a foothold. Too much time is spent worrying about Murchison's political incorrectness, and the editors refuse point-blank to discuss his "supposed imperialism" (as if that were ever in doubt), although this aspect of his career has recently excited more interest than any other.

The bibliography has some striking omissions, in particular James A. Secord's

ground-breaking article “King of Siluria” in *Victorian Studies* 25 (1981–82) and his subsequent book *Controversy in Victorian Geology* (Princeton University Press, 1986), and Robert Stafford’s book *Scientist of Empire* (Cambridge University Press, 1990). Along with the work of Martin Rudwick, these studies provide the historical context lacking in Collie and Diemer’s introduction.

But it is by the text itself that such a book stands or falls. For making this important narrative available in such an attractive and (for its quality) inexpensive format, the editors are to be congratulated. ■

Ralph O'Connor is a fellow of St John's College, University of Cambridge, Cambridge CB2 1TP, UK. He is writing *The Greatest Show on Earth: Staging Prehistoric Worlds for the British Public, 1802–1856*.

Theatre

Playing dirty

Calculus

Written by Carl Djerassi
Performed at the New End Theatre, London,
until 28 August 2004

Philip Ball

We don't generally expect our geniuses to be genial, but few are as downright misanthropic as the version of Isaac Newton now in vogue in science histories. He is the villain in Lisa Jardine's *The Curious Life of Robert Hooke* (HarperCollins, 2003) who

tried to more or less erase the hapless Hooke from history. He is an elusive, forbidding presence in James Gleick's popular recent biography *Isaac Newton* (Pantheon, 2003). And Stephen Hawking, who now occupies Newton's chair at Cambridge, has accused him of being vindictive, arrogant and prone to petty arguments.

In Carl Djerassi's new play, *Calculus*, Newton appears just once — as a monster who browbeats Queen Anne's physician John Arbuthnot with furious silences. But this may not be the real Newton after all: his appearance features in a play within a play, written by architect John Vanbrugh and theatre manager Colley Cibber to reveal the 'truth' about Newton's dealings with the Royal Society in his dispute with Gottfried Wilhelm Leibniz.

Djerassi has used this device before in his play *Oxygen* (2001), which was also about priority disputes in the history of science. It is a useful way of dealing with the complexities of the issues. It allows Cibber to anticipate the audience's dismay (and indeed I sensed such a response) at having to hear about the calculus, the mathematical technique devised independently by Newton and Leibniz. It also lets Cibber and Vanbrugh impersonate the two protagonists as they argue over the details. The problem is that Arbuthnot's final revelation to Cibber about the source of his play's material forces us to doubt whether what we have just witnessed is any reflection of what really transpired when the Royal Society was called upon to adjudicate in the dispute.

Ultimately, however, *Calculus* struggles with the fact that there is just not quite enough at stake here to sustain the drama. The dialogue is crisp and the staging attractive, but the central event — Newton's manipulation of the Royal Society — is not inherently theatrical and does not generate sufficient tension. (I did, however, enjoy the portrayal of the eminent French mathematician Abraham De Moivre as a gluttonous reprobate.)

Newton called his version of the differential calculus 'fluxions'. He developed it in the 1660s, when he was in his twenties, but like so much of his work, he declined to publish it until it seeped out in the first version of the *Principia* in 1687. But when Leibniz visited London in 1673, it became clear to Henry Oldenburg, secretary of the Royal Society, that trouble lay ahead, for the German mathematician had devised a similar method. Newton was persuaded to write to Leibniz in 1676 to point out his prior work, and although the two men corresponded amicably enough for the next year, they reached no agreement. Leibniz published his own version of the calculus in 1684, but it was not until 1699 that the real arguments began. That was when Newton's friend Fatio de Duillier more or less accused Leibniz of plagiarism.

Djerassi wisely avoids trying to resolve this dispute. Rather, his play centres on the deliberations of a Royal Society committee appointed in 1712 to pronounce on the priority issue. Newton was the society's president at the time, so it was no surprise that most of the committee's 11 members were his supporters. But he took no chances, drafting the committee's report himself (it stated that “we reckon Mr Newton the first inventor” of the calculus) and simply presenting it to them for signing. They duly approved it, and Newton stepped in again to take care of the printing. Leibniz knew nothing of the process until the declaration was distributed to scholars all over Europe.

It was surely one of Newton's most deplorable acts. Djerassi looks at the ways that the committee members wrestled with their consciences, and at how considerations of nationalism, religion and personal politics coloured their behaviour. But in the end, the play's broader question is about scientific reputation: is it better to reveal science's shameful secrets or to preserve its appearance of authority? Djerassi says that he was himself accused of “washing dirty lab coats in public” after writing his earlier fictional books on scientific research. Mercifully we no longer seem compelled to sanitize the history and sociology of science in the way that David Brewster did with his 1831 hagiography of Newton, in which he meticulously excised all mention of the great man's interests in alchemy and astrology. ■

Philip Ball is a consultant editor for *Nature*.



A play within a play: Colley Cibber and John Vanbrugh make a point about Newton in *Calculus*.

On the move

Heterogeneous catalysis: just as for enzymes, flexibility and mobility are emerging as key features of catalytically active metal surfaces.

Gabor A. Somorjai

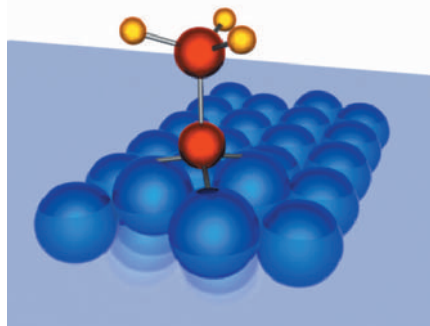
Catalysts have the remarkable property of facilitating a chemical reaction repeatedly without being consumed. In principle (if not in practice), a catalyst can function as long as reactants are available. Enzymes — which catalyse all aspects of cell metabolism — are the supreme masters of this art. We have used them from earliest times for leavening bread, curdling cheese and brewing beer.

In the twentieth century, a seemingly very different type of catalysis — heterogeneous catalysis, which uses small metal particles supported on solid surfaces — became the foundation for much of the chemical industry. It plays a central role in generating the feedstocks for making the synthetic materials that we use every day, from fuels to fertilizers. New experimental techniques have brought fresh insights into this form of catalysis, and it now seems that there are more similarities between enzymes and heterogeneous catalysts than initially meets the eye.

Heterogeneous catalysis usually involves the interaction between a reactant and the catalytic surface of small metal particles. This in turn activates chemical bonds within the reactant, resulting in their dissociation or rearrangement. For decades, studies of heterogeneous catalysts were largely empirical, involving little understanding at the molecular level. Theories of how they function emphasized the need for complementarity between reactant structure and the geometrical arrangement of atoms in the static catalyst surface. Now, more advanced methods of observation have allowed us to see these molecules during reactions in considerably more detail, and have revealed that this static view of catalysis is, in fact, wrong.

The advent of ultrahigh-vacuum surface science in the 1960s provided an entrée into the world of adsorbed molecules. However, it is only recently that modern techniques such as high-pressure scanning tunnelling microscopy have allowed experiments to be done under realistic (high pressure and temperature) conditions.

Exposure to reactants results in local restructuring of the catalytic surface. The formation of bonds between the metal atoms of the catalyst and the adsorbed molecules produces heat, providing energy to loosen the bonds between the metal atom and its neighbours. This allows it to move from its original position and thus strengthens the



All change: formation of the catalytic intermediate ethylidyne causes catalyst surface restructuring.

chemical bond between the adsorbed molecule and the metal. The fewer neighbours a surface metal atom has, and the stronger the bonding to the adsorbed molecules, the greater is the degree of restructuring that occurs. Also, it is not always the original reactant that bonds with the metal catalyst; it can be a derivative if the metal–adsorbate complex can form stronger bonds this way.

The formation of strong chemical bonds at these active metal sites intuitively suggests that such sites will be blocked and prevented from participating in new reactions. But in fact the very same sites that are most active in bond activation are also catalytically the most active, and remain so. They can carry out bond activation repeatedly.

The reason for the continued catalytic activity is that it is not only the metal atoms that are on the move. Strongly adsorbed reactants on the metal surface are also mobile. In a model reaction (the hydrogenation of ethene to form ethane) used to understand the industrially important hydrogenation process, an important derivative species, ethylidyne, retains the ability to move over the surface despite being fairly strongly bound. The ethylidyne is not part of the catalytic process, but remodels active sites and then moves on, allowing ethene, which is adsorbed only weakly, to bind briefly and become rapidly hydrogenated. Impeding the mobility of ethylidyne slows the catalytic reaction. For example, if carbon monoxide is adsorbed on the catalyst surface, ordered, immobile surface structures containing both ethylidyne and carbon monoxide are formed which prevent mobility and block adsorption sites, ‘poisoning’ the catalytic reaction.

Thus, everything on the surface must move for catalysis to occur. The metal atoms move to restructure the surface as the ethylidyne forms; ethylidyne moves, thereby

opening up active sites that allow new ethene molecules to adsorb weakly and hydrogenate rapidly. Without mobility there is no heterogeneous catalysis.

Flexibility in restructuring the sites where bond activation occurs is well known to those working on enzyme catalysts. Emil Fischer’s static lock-and-key model for the interaction of an enzyme with its substrate was superseded by the induced-fit dynamic model of Daniel Koshland. In this model, chemisorption of the reacting molecule induces restructuring of the active enzyme, which results in structural and electronic alterations to both. Substrate binding and product release require the restructuring of both the active site and surrounding parts of the enzyme. Meanwhile, chemical-bond activation and catalytic turnover probably require both local restructuring and mobility of adsorbed species to and from the active sites.

The similarity in the behaviour of heterogeneous and enzyme catalysts at the molecular level could lead to the discovery of new selective catalysts if these two presently diverse fields were studied together. However, the natural habitat of enzymes is water at or near room temperature, whereas heterogeneous catalysts usually operate at high temperatures and in the presence of gaseous or liquid reactants at high pressures. Bridging the chasm between these diverse conditions and the vastly different chemical natures of the biological and industrial ‘catalyst worlds’ will be a real challenge. But if successful, such an approach could lead to new generations of selective catalysts that have both the high turnover, poison resistance and thermal stability of present-day heterogeneous systems and the selectivity for complex chemical rearrangements of enzymes. In short, it could help us to realize the goal of ‘green chemistry’ — producing desired product molecules while minimizing chemical waste. ■

Gabor A. Somorjai is in the Department of Chemistry, University of California, Berkeley, and the Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA.

FURTHER READING

Somorjai, G. A. *Introduction to Surface Chemistry and Catalysis* (Wiley, New York, 1994).

Thomas, J. M. & Thomas, W. J. *Principles and Practice of Heterogeneous Catalysis* (Wiley-VCH, Berlin, 1997).

Tsou, C.-L. *Science* **262**, 380–381 (1993).

Quantizing the classical cat

Ian Stewart

A mathematical analysis of a pendulum system reveals the relevance to quantum systems of the classical concept of 'monodromy' — why a falling cat always lands the right way up.

A central problem in modern physics is to find effective methods for quantizing classical dynamical systems — modifying the classical equations to incorporate the effects of quantum mechanics. One of the main obstacles is the disparity between the linearity of quantum theory and the nonlinearity of classical dynamics. Taking a big step forward, R. H. Cushman *et al.* have analysed a quantum version of the spring pendulum, whose resonant state was first discussed by Enrico Fermi and which is a standard model for the carbon dioxide molecule (*Phys. Rev. Lett.* **93**, 024302; 2004).

Cushman *et al.* show that when this system is quantized, the allowed states, or eigenstates, fail to form a perfect lattice, contrary to simpler examples. Instead, the lattice has a defect, a point at which the regular lattice structure is destroyed. They show that this defect can be understood in terms of an important classical phenomenon known as monodromy. A quantum-mechanical cliché is Schrödinger's cat, whose role is to dramatize the superposition of quantum states by being both alive and dead. Classical mechanics now introduces a second cat, which dramatizes monodromy through its ability always to land on its feet (Fig. 1). The work affords important new insights into the general problem of quantization, as well as being a beautiful example of the relation between nonlinear dynamics and quantum theory.

The underlying classical model here is the swing–spring, a mass suspended from a fixed point by a spring (Fig. 2a, overleaf). The spring is free to swing like a pendulum in any vertical plane through the fixed point, and it can also oscillate along its length by expanding and contracting. The Fermi resonance occurs when the spring frequency is twice the swing frequency. The same resonance occurs in a simplified model of the two main classical vibrational modes of the carbon dioxide molecule (Fig. 2b), and the first mathematical analysis of the swing–spring was inspired by this model.

Using a modern technique of analysis known as reduction, which exploits the rotational symmetry of a system, Cushman *et al.* show that this particular resonance has a curious implication, which manifests itself physically as a switching phenomenon. Start with the spring oscillating vertically but in

a slightly unstable state. The vertical 'spring mode' motion quickly becomes a 'swing mode' oscillation, just like a clock pendulum swinging in some vertical plane. However, this swing state is transient and the system returns once more to its spring mode, then back to a swing mode, and so on indefinitely. The surprise is that the successive planes in which it swings are different at each stage. Moreover, the angle through which the swing plane turns, from one occurrence to the next, depends sensitively on the amplitude of the original spring mode.

The apparent paradox here is that the initial state has zero angular momentum — the net spin about the vertical axis is zero. Yet the swing state rotates from one instance to the next. Analogously, a falling cat that starts upside down has no angular momentum about its own longitudinal axis, yet it can invert itself, apparently spinning about that axis. The resolution of the paradox, for a cat, is that the animal changes its shape by moving its paws and tail in a particular way. At each stage of the motion, angular momentum remains zero and is thus conserved, but the overall effect of the shape changes is to invert the cat. The final upright state also has zero angular momentum, so there is no contradiction of conservation. This effect is known as the 'geometric phase', or monodromy, and is important in many areas of physics and mathematics.

The central topic of the paper is this: how does monodromy show up when the system is quantized? The answer, obtained in the specific context of the carbon dioxide molecule, is both elegant and remarkable.

A molecule of carbon dioxide can be modelled classically as a central carbon atom, attached symmetrically by identical springs to two oxygen atoms, with the springs inclined at an obtuse angle (Fig. 2b). The molecule has three main vibrational modes. The two most important modes are symmetric stretching, where both springs change their lengths in synchrony, and bending, where the angle between the two springs oscillates. These modes are analogous to the spring and swing modes of a swing–spring. The third main mode, asymmetric stretching, occurs when the two springs oscillate out of phase with each other, and it can be removed from consideration by averaging over a vibrational cycle. The result is a 'reduced Hamiltonian', or energy function,

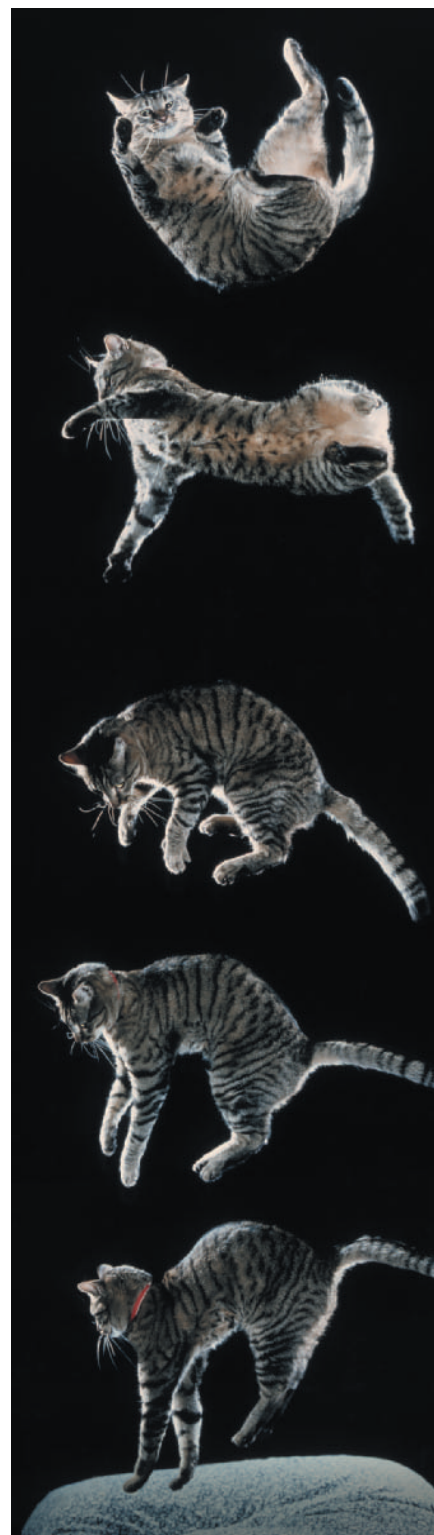


Figure 1 The classical cat always lands on its feet.



100 YEARS AGO

The result of this inquiry is to prove the existence of a small number of more or less isolated hereditary centres, round which a large part of the total ability of the nation is clustered, with a closeness that rapidly diminishes as the distance of kinship from its centre increases. The materials are derived from the replies to a circular which I sent with a blank schedule, to all fellows of the Royal Society, asking for the names and achievements of their "noteworthy" kinsfolk in each degree of near kinship as specified in the schedule. Noteworthiness was defined as including any success that was, in the opinion of the sender, at least equal in its way to that in which the honour of a fellowship of the Royal Society is held by scientific men. Returns are still dropping in, and now exceed two hundred. They continue to be very acceptable, but I judged it best to content myself with the number received up to a date when I could conveniently work at them, and to publish the preliminary results without delay... the experience gained through this inquiry has strongly confirmed an opinion expressed in my lecture on Eugenics before the Sociological Society... that it would be both feasible and advantageous to make a register of gifted families.

Francis Galton

From *Nature* 11 August 1904.

50 YEARS AGO

The chromosomes of *Mus musculus* have a high chiasma frequency, and for this reason very loose linkages are to be expected. Many of the problems of linkage and independence in this species may therefore have to be solved by cytogenetic methods rather than the breeding techniques of formal genetics. Among them is the question whether linkage group VII is carried in the pairing segment of the sex chromosome... With the object of obtaining evidence on questions such as this we have induced a number of translocations in the mouse, using X-rays, and have identified linkage groups in eleven of them... Translocation 78 thus offers a means of settling the question whether linkage group VII is sex-linked. The translocation and the sex bivalent should be cytologically recognizable in primary spermatocytes; it should therefore be possible to establish their chromosomal independence or interdependence.

T. C. Carter, Mary F. Lyon & Rita J. S. Phillips

From *Nature* 14 August 1954.

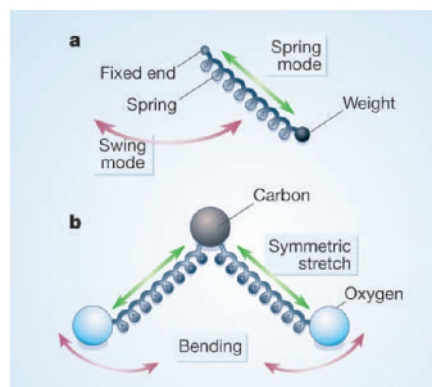


Figure 2 The swing-spring. The swing-spring can stretch like a spring and swing like a pendulum (a), and can be used as a simple model for the carbon dioxide molecule (b).

which is simpler than the exact Hamiltonian but is still a good model.

The quantum energy-momentum lattice of the molecule consists of the eigenstates of this Hamiltonian, that is, the pure vibrational modes. For a fixed energy, these modes correspond to two classical 'constants of motion'—angular momentum and a quantity related to the rotational symmetry. The eigenstates can be characterized by two quantum numbers, which are integers, so these eigenstates form a regular planar lattice like a chessboard.

However, there is an extra quantum number, related to another classical variable, called the 'action'. The new phenomenon

here is that, because of monodromy, the action is defined only locally and cannot be consistently extended across the entire lattice. For fixed quantum numbers in the lattice, this additional quantum number can take on infinitely many values, at equally spaced points at right angles to the chessboard. The simplest structure of this kind is a three-dimensional cubic lattice—an infinite stack of chessboards, vertically above each other. Monodromy implies that the totality of all sets of quantum numbers does not form a cubic lattice. Instead, it has a single topological defect where the regularity of the lattice structure breaks down.

This analysis is important because it suggests, and supports, a general principle. The most significant features of the quantum-mechanical description of a classical system occur at its singularities. The singularities introduce defects into the ensemble of quantum eigenstates, but they also organize the structure of those defects. Everywhere else, quantization works just as in previous, simpler examples. The authors suggest several directions for future progress, mostly to develop the growing use of nonlinear dynamics in the understanding of quantization. But the most tantalizing is the possibility of detecting quantum monodromy experimentally. Maybe we will soon be able to see how Schrödinger's cat turns itself upside down. ■

Ian Stewart is at the Mathematics Institute, University of Warwick, Coventry CV4 7AL, UK. e-mail: ins@maths.warwick.ac.uk

Cognitive science

Rank inferred by reason

Sara J. Shettleworth

Pinyon jays seem to work out how to behave towards an unfamiliar jay by watching it in encounters with members of their own flock. The findings provide clues about how cognition evolved in social animals.

Susan is taller than Billy. Peter is taller than Susan. Who is taller, Billy or Peter? Knowledge about pairs of objects linked by relationships such as 'taller' or 'stronger' permits conclusions to be drawn about novel pairs (here, Billy and Peter)—a process known as transitive inference. Monkeys, rats and some birds can solve transitive-inference tasks in the laboratory¹, but why would this ability evolve? A plausible answer is that transitive inference is an evolutionary adaptation in certain kinds of social group. For example, suppose I know from bitter experience that Bob always beats me in contests (that is, he dominates me). I now observe some new individual, Andy, dominating Bob. If I reason, "Andy dominates Bob, and Bob dominates me, therefore Andy will dominate me", I can avoid fights by deferring to Andy when we meet. But there

has been no well-controlled evidence that animals actually use transitive inference in social situations. In the study reported on page 778 of this issue, Paz-y-Miño and colleagues² provide this.

In effect, the authors staged the Andy-and-Bob scenario using pinyon jays (*Gymnorhinus cyanocephalus*; Fig. 1), a highly social member of the crow family. These birds live in large, permanent flocks with clear pecking orders. Paz-y-Miño and colleagues created groups of captive pinyon jays that were previously unknown to each other, and allowed stable dominance relationships to develop in each group. Then jays from each group were allowed to observe individuals from other groups interacting over a peanut, and later interacted with some of those same birds. In the experiment, the observer saw a relatively dominant bird from



Figure 1 Pinyon jays — watching and learning.

its own group ('Bob' in our scenario) losing encounters with a stranger from another group ('Andy'). To ensure that the stranger would not be seen only to dominate others, the observer also watched the same stranger losing contests with another stranger. As a control, other observer birds watched a stranger both winning and losing to members of the stranger's own group, an experience that should give observers no information on whether the stranger will be able to dominate them.

If pinyon jays infer social status transitively, observers should behave more submissively in their first encounter with the stranger in the experimental condition than in the control — and this is what Paz-y-Miño and colleagues² observed. To act in this way, the observers must first have identified both the individuals they watched and their roles in the observed encounter, and then retained this information for later use. Thus, their behaviour implies a more complex set of cognitive skills than that implied by another recent report of birds' sensitivity to a generic social relation (whether mated or not) between unfamiliar individuals present³.

It is noteworthy that the animals in this study are birds. The idea that the demands of a complex social life might drive the evolution of cognition was originally proposed to explain the apparently high intelligence of monkeys and apes⁴. However, other mammals, such as hyenas and elephants, and some birds also form long-lasting groups of identifiable individuals with differentiated social roles. Long-term field studies indicate that social complexity has shaped cognition in a similar way across species^{5,6}, but in field work it can be difficult to know all the animals' relevant experiences. The experimental approach of Paz-y-Miño and colleagues² will serve as a model of how manipulating

the experiences of captive animals can provide firm conclusions.

If transitive inference is used in social life, species with more complex societies should be better at it. Laboratory studies using an abstract transitive-inference task provide some support for this idea. First, animals are taught the relative reward value of five or more pairs of colours⁷ — red is better than green, green is better than blue, and so on. Then they are tested with novel pairings. Monkeys behave in such tests as though they are reasoning by transitive inference, whereas pigeons do not¹. But this finding need not reflect a difference in sociality between monkeys and pigeons because they differ in so many other ways. More relevant is a recent comparison of pinyon jays with the closely related, but less social, western scrub jay⁸. Pinyon jays perform more like monkeys in the abstract task than do scrub jays.

Members of the crow family, such as New Caledonian crows⁹, western scrub jays¹⁰ and ravens, feature in recent reports of remarkable cognitive abilities including making and using tools, episodic-like memory and forms of social learning. Each of these abilities is thought to reflect a species-specific adaptation, for example to extract prey or retrieve stored food. Comparisons of different species

in similar tasks are necessary to test whether such abilities are in fact associated with the ecological factors supposed to select for them. Nevertheless, such findings underline the importance of viewing animal intelligence as consisting partly of species-specific adaptations, rather than being an entirely undifferentiated 'general intelligence'. So, comparative psychology can provide experimental support for the idea that cognitive modules have evolved for solving different kinds of task — an idea so popular, but often untestable, in human evolutionary psychology⁵.

Sara J. Shettleworth is in the Departments of Psychology and Zoology, University of Toronto, Toronto, Ontario M5S 3G3, Canada.

e-mail: shettle@psych.utoronto.ca

1. Shettleworth, S. J. *Cognition, Evolution, and Behavior* (Oxford Univ. Press, 1998).
2. Paz-y-Miño C. G., Bond, A. B., Kamil, A. C. & Balda, R. P. *Nature* **430**, 778–781 (2004).
3. Vignal, C., Mathevon, N. & Mottin, S. *Nature* **430**, 448–451 (2004).
4. Jolly, A. *Science* **153**, 501–506 (1966).
5. Kamil, A. C. *Trends Cogn. Sci.* **8**, 195–197 (2004).
6. de Waal, F. B. M. & Tyack, P. L. (eds) *Animal Social Complexity* (Harvard Univ. Press, Cambridge, 2003).
7. McGonigle, B. O. & Chalmers, M. *Nature* **267**, 694–696 (1977).
8. Bond, A. B., Kamil, A. C. & Balda, R. P. *Anim. Behav.* **65**, 479–487 (2003).
9. Weir, A. A. S., Chappell, J. & Kacelnik, A. *Science* **297**, 981 (2002).
10. Clayton, N. S. & Dickinson, A. *Nature* **395**, 272–274 (1998).

Human genetics

An expression of interest

Nancy J. Cox

The baseline level of gene expression varies from person to person, but how is this determined genetically? The answer may improve our understanding of complex traits, including some genetic diseases.

On page 743 of this issue, Morley and colleagues¹ report how they successfully determined the genomic locations of genetic determinants that influence the variation in levels of gene expression between people. How was this accomplished, and why is it important?

In searching for the genetic basis of a measurable trait, or phenotype — height or eye colour, for example — geneticists start by identifying people with variation in that trait. In the simplest situations, the variation is traced to a single gene, and particular variations in the sequence of that gene are shown to determine the observed phenotypic variation.

A less tangible, but no less significant, trait is the baseline level of gene expression. Much of our progress so far in understanding the basis of rare genetic diseases has come through identifying changes in gene sequence that change the nature of the encoded protein, so that it is insufficiently or inappropriately functional. But changes in the amount of protein produced might also

affect the health of an organism. This amount is determined in part by gene expression — or how much messenger RNA (mRNA) is transcribed from the relevant DNA sequence — and the relative abundances of mRNA for many thousands of genes can now be routinely assessed for any accessible tissue. But gene expression is generally regulated by DNA regions outside the parts of genes that actually encode proteins, and there is much that is not yet understood about this process. In particular, little is known about how variation in DNA sequences might affect the variability in baseline levels of gene expression among individuals — the topic of Morley and colleagues' investigations¹.

Thus, their studies stand squarely at the crossroads of some of the most challenging questions in human genetics. Among the most pressing of these questions are: how can we identify regulatory elements? Does variation within regulatory elements have the same frequency spectrum as the variation that occurs within genes and

demonstrably affects the function of their encoded proteins? And to uncover the genetic variation affecting complex traits, will it be more efficient to survey variation within genes or within regulatory elements? The answers to these questions will have tremendous implications for how geneticists design today's studies of candidate genes for complex traits. They will also influence how we target investments in genomics infrastructure to have the most positive impact on future studies of genetic determinants for complex traits, including common human diseases. So far, researchers have had to make design and investment decisions with a relatively rudimentary understanding of the nature, and even the extent, of the genetic variation that underlies variability in human gene expression.

Previous studies have shown that variation in gene-expression phenotypes in 'immortalized' human B cells (a type of immune cell that can replicate indefinitely in culture) is heritable². Moreover, the genetic variation affecting similar phenotypes in a variety of organisms, including yeast and mice, can be located through a genetic tool known as linkage mapping^{3,4}, which involves following the trait and large numbers of genetic markers through pedigrees. Now, Morley *et al.*¹ show that this tool can also be used to pinpoint the genetic determinants of variation in gene expression in human cells — and on a genome-wide scale.

The authors started by ascertaining the baseline expression level of roughly 8,500 genes that are active in immortalized human B cells from 14 families. They then focused on the subset of genes (3,554) with more variation in expression level between unrelated individuals than between replicate samples within individuals. They used a recent modification⁵ of a classic approach⁶ to map this variation to specific chromosomal regions. The authors found that variation in expression of 984 of these genes showed highly significant evidence for linkage to one or more regions of the genome. Some of these regions work in *cis* (the determinant is adjacent to the gene it regulates), but most seem to work in *trans* (the determinant is distant from the gene it regulates, often on a different chromosome). The success of Morley and colleagues' linkage-mapping studies in localizing regulatory regions was reinforced by the results of their association studies of some of the *cis* regulators. These showed that genetic variation in or very near these genes strongly predicted the phenotypic variability in gene expression.

The investigators characterize their results as indicating that the human gene-expression phenotype is "a trait like many others... amenable to genetic analysis". To veterans of linkage-mapping studies on complex human phenotypes, this may seem to be an understatement on a par with Wat-

son and Crick's "it has not escaped our notice..."⁷. Indeed, these gene-expression phenotypes seem to be particularly amenable to genetic analysis, which bodes well for the future studies that will build on these results. Part of the success of Morley and colleagues' studies is undoubtedly attributable to their design. By focusing first on genes that show greater variance in expression levels between unrelated individuals than between replicates from the same individual, they almost certainly improved the likelihood of successfully locating the underlying genetic variation. And, by using cells from 14 previously studied pedigrees, they were able to adopt existing data on genetic variation for the linkage studies.

There are limitations to the studies. As noted above, mRNA levels are only part of what determines protein levels. Also, it is unknown whether the patterns of *cis* and *trans* regulation deduced from the linkage mapping are general, or whether they vary at different stages of development and in different tissues. This issue may be more difficult to address through direct linkage mapping, given the limited number of tissues that can be easily tested in family members. Once genetic variation has been reliably associated with human expression phenotypes, however, it is more straightforward to determine whether there is important variability in the observed associations, either temporally or by tissue.

The promise of these studies is equally clear. Among the most interesting of Morley and colleagues' findings¹ are the regions that the authors propose as master

regulators of baseline levels of gene expression. These master regulators are locations that show strong evidence of linkage to the expression of many genes, some of which have highly correlated levels of expression. Such regions are particularly attractive targets for follow-up studies designed to identify the precise sequence variations underlying the linkage signals. This should in turn give much-needed insight into mechanisms of regulation that remain poorly characterized.

Moreover, these studies should help researchers to identify and validate regulatory networks that can be used in bioinformatics approaches to prioritize genotype-phenotype studies. Finally, further studies of the strong linkage signals and corroborative associations that have already been achieved with human gene expression might provide welcome insight into the generic challenges of identifying the genetic variation underlying other complex phenotypes. It seems a safe bet that interest in the genetic basis of variation in human gene expression will be a stable phenotype — with a high level of expression — for some years to come. ■

Nancy J. Cox is in the Department of Human Genetics, University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA.
e-mail: ncox@genetics.uchicago.edu

1. Morley, M. *et al.* *Nature* **430**, 743–747 (2004).
2. Cheung, V. G. *et al.* *Nature Genet.* **33**, 422–425 (2003).
3. Brem, R. B. *et al.* *Science* **296**, 752–755 (2002).
4. Schadt, E. E. *et al.* *Nature* **422**, 297–302 (2003).
5. Shete, S. *et al.* *Hum. Hered.* **55**, 79–85 (2003).
6. Haseman, J. K. & Elston, R. C. *Behav. Genet.* **2**, 3–19 (1972).
7. Watson, J. D. & Crick, F. H. C. *Nature* **171**, 737–738 (1953).

Cell biology

Barbed ends rule

Dorothy A. Schafer

To explore their surroundings, cells use probes of various shapes. Whether the probes are broad and flat, or long and thin, seems to be regulated by proteins at the growing ends of actin filaments.

Cells reach out with various structures as they crawl and interact with their environment. Broad, flat protrusions called lamellipodia surge forward and adhere to surfaces, allowing cells to gain traction and move. Long, thin protrusions called filopodia extend several micrometres ahead of the cells, appearing to explore extracellular surfaces, sense guiding cues and direct the rest of the cell. Extension of both lamellipodia and filopodia is powered by polymerization of the structural protein actin into filaments. As the filaments lengthen, the fast-growing 'barbed' ends push the cell's boundary membrane outward. Lamellipodia protrude when the actin filaments at the leading edge are short and highly

branched; by contrast, filopodia arise when the filaments are long, unbranched, and arranged in tight, parallel bundles. But given the similarities in the actin polymerization machinery that drives the two types of protrusion, what determines whether lamellipodia or filopodia form? Marisan Mejillano and colleagues, writing in *Cell*¹, report that the answer lies with proteins that control actin polymerization at the filament barbed ends.

Two models have been proposed to explain the formation of these actin-based protrusions — one dealing with lamellipodia and one with filopodia (Fig. 1). The dendritic-nucleation model² suggests that lamellipodia arise when new actin filaments form as branches on existing actin filaments,

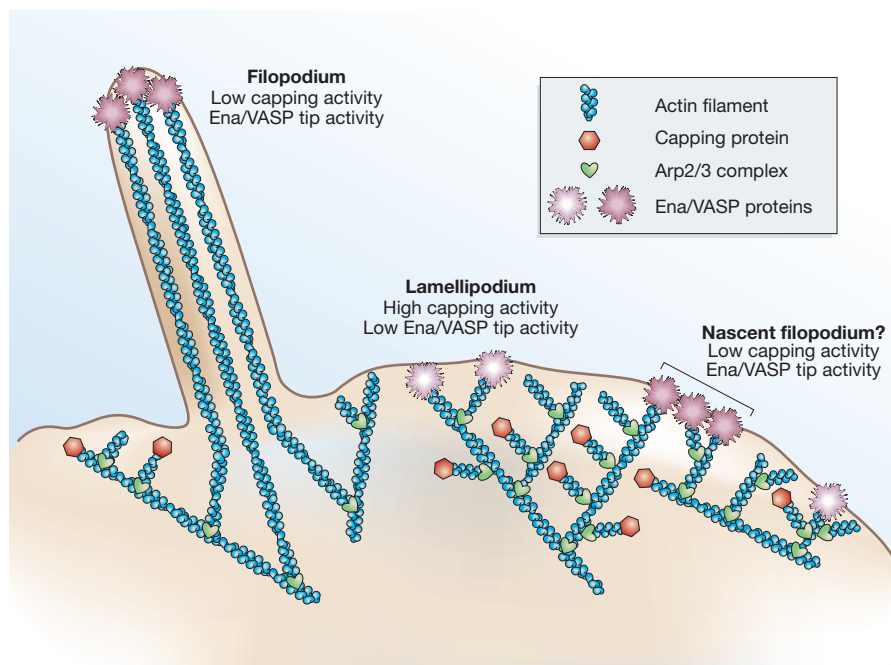


Figure 1 Model for how the growing ends of actin filaments regulate cellular protrusions. Actin filaments polymerize at their fast-growing 'barbed' ends near the plasma membrane. They stop growing when they are capped by capping protein. If barbed-end capping activity is high, filaments get capped, stop growing and remain short. This favours an actin filament network that is most efficient for lamellipodium expansion. Filopodia result when filaments of the lamellipodial actin network are recruited by a filopodial tip complex that includes Ena/VASP proteins; filaments grow long because Ena/VASP proteins also inhibit capping locally.

under the direction of a protein complex called Arp2/3. The filaments push the membrane outward as they elongate by polymerization, until the growing ends are 'capped'. Capping protein binds tightly to the barbed ends, blocking filament growth³. The result is a lamellipodial actin network of short, stiff, branched filaments. According to the convergent-elongation model⁴ of filopodium development, a 'tip complex' of proteins recruits filaments of the lamellipodial network to become the filopodium. The tip complex includes Ena/VASP proteins, which encourage the growth of long, unbranched filaments by inhibiting the capping process⁵. Ena/VASP proteins might also recruit other proteins to further stabilize and organize the actin filaments into bundles as the filopodia lengthen. Both models suggest that whether or not capping occurs at the growing end of actin filaments is key to which protrusive structure forms.

So Mejillano and colleagues¹ studied the effects of reducing the amount of capping protein on the formation of protrusions, using cells that normally form both types. They found that depleting the cells of capping protein dramatically changes both lamellipodia and filopodia. Most striking was a massive induction of filopodium-like protrusions all over the cell surface. The protrusions were bona fide filopodia: they exhibited the motile behaviour of filopodia, contained parallel bundles of long actin filaments, and recruited two molecular

hallmarks of filopodia, Ena/VASP proteins and fascin, a protein that helps to bundle the actin filaments.

In contrast to the explosion of filopodia, the normally smooth, coordinated lamellipodial protrusions became slow and erratic following the reduction in levels of capping protein. This behaviour would be expected if the cellular pool of actin monomers is used up in fuelling the rampant polymerization of filopodial actin filaments, rather than contributing to the lamellipodia. In addition, lamellipodial filamentous actin and Arp2/3 complex, which are normally enriched at the cell edge, were lost. The cytoplasm further away from the cell periphery became dense with actin filaments and Arp2/3 complex, suggesting that the machinery that disassembles actin filaments away from the cell edge is disrupted when barbed-end capping activity is low.

So, lowering barbed-end capping activity seems to tweak the actin machinery at the cell's leading edge towards polymerization of actin filaments in filopodia. Densely packed actin filaments that are inefficient at pushing the membrane forward thus replace the lamellipodial actin network. When capping protein was reinstated in the depleted cells, excess filopodia disappeared and lamellipodia were restored, indicating that capping protein is entirely responsible for the effect of barbed-end capping activity on cellular protrusions.

But why do filopodia, and not lamellipodia,

grow explosively when barbed ends remain uncapped? Mejillano *et al.*¹ discovered that Ena/VASP proteins also influence which type of protrusion the cell extends. Because Ena/VASP proteins inhibit barbed-end capping activity *in vitro*⁵, the team predicted that reducing capping protein in cells lacking these capping antagonists would further enhance filopodia at the expense of lamellipodia. But when they tried this experiment, no filopodia formed; instead, ruffles were produced along the cell edge. Most probably, these result when actin filaments at the membrane become so long and flexible that they buckle and collapse rearward as they fail to sustain a pushing force. So loss of barbed-end capping activity clearly enhances actin polymerization at the cell periphery, but is not sufficient to specify whether filopodia or lamellipodia protrude when Ena/VASP proteins are absent. However, when the Ena/VASP protein known as Mena was restored in previously depleted cells, the formation of abundant filopodia returned, highlighting a special role for Ena/VASP proteins in initiating filopodia, even if barbed-end capping activity is low.

Mejillano and colleagues' findings provide evidence that supports key features of the two models of actin-filament-based protrusion. First, as proposed in the dendritic-nucleation hypothesis², regulated barbed-end capping is essential to maintain productive lamellipodia. Lamellipodia extend persistently when barbed ends are capped because the unpolymerized actin pool is maintained, and actin filaments of the network remain short and, hence, sufficiently stiff to push the membrane forward. Second, as predicted by the convergent-elongation hypothesis⁴, the formation of filopodia requires Ena/VASP proteins. These proteins seem to be the glue of the filopodial tip complex that recruits lamellipodial actin filaments and promotes their growth to form nascent filopodia.

So, when it comes to actin-filament-based protrusions, regulated polymerization at filament barbed ends is essential, and both capping protein and Ena/VASP proteins contribute by regulating barbed-end capping activity. Ena/VASP proteins are also essential for a functional filopodial tip complex. With Mejillano and colleagues' findings, therefore, we have a clearer picture of how barbed ends point the way. ■

Dorothy A. Schafer is in the Departments of Biology and Cell Biology, University of Virginia, Charlottesville, Virginia 22903, USA.

e-mail: das9w@virginia.edu

1. Mejillano, M. R. *et al.* *Cell* **118**, 363–373 (2004).

2. Pollard, T. P., Blanchoin, L. & Mullins, R. D. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 545–576 (2000).

3. Cooper, J. A. & Schafer, D. A. *Curr. Opin. Cell Biol.* **12**, 97–103 (2000).

4. Svitkina, T. M. *et al.* *J. Cell Biol.* **160**, 409–421 (2003).

5. Bear, J. E. *et al.* *Cell* **109**, 509–521 (2002).

Biological chemistry

The making of Moco

William N. Hunter

Nature exploits the unique chemistry of molybdenum in many reactions. Structures of the enzyme Cnx1 reveal unexpected mechanisms for slotting the metal ion into its reactive position in the cofactor Moco.

From bacteria to mammals, molybdenum (Mo) is crucial for survival. Enzymes that take advantage of the distinctive redox chemistry of this metal are involved in numerous metabolic reactions in the carbon, nitrogen and sulphur cycles^{1,2}. To be available for these reactions, however, water-soluble molybdate anions from the environment must be bound, transported into cells and manipulated to place them in the correct context to exploit their chemical properties. This complex manoeuvring of Mo involves an ancient and highly conserved biosynthetic pathway, and a key enzyme of the pathway in plants is called Cnx1. On page 803 of this issue, Kuper *et al.*³ report two crystal structures of the Mo-binding domain of Cnx1. Unexpectedly, the structures suggest a new step in the pathway, involving an adenosine metabolite, and they implicate copper ions in the process as well — two fortuitous and intriguing observations.

The vast majority of enzymes that make use of molybdenum (termed molybdoenzymes) do so through the Mo cofactor — Moco for short. This cofactor contains a single Mo coordinated to an organic molecule called molybdopterin. The biosynthesis of Moco is highly conserved from bacteria through to mammals and involves four stages (Fig. 1): first, conversion of a guanine nucleotide into a derivative termed precursor Z; second, conversion of precursor Z into molybdopterin; third, binding of Mo by molybdopterin, producing Moco; and fourth, attachment of a nucleotide moiety to Moco, forming molybdopterin guanine dinucleotide. In eukaryotes, Moco is considered the active form of the cofactor, but most bacterial enzymes require the molybdopterin guanine dinucleotide to make them functionally active.

How Moco is then inserted into molybdoenzymes is not yet understood because of the intrinsic instability of the chemical species concerned, but the intermediates and cofactor probably remain bound to other proteins throughout the biosynthetic process until the final incorporation of Moco to fully constitute the molybdoenzymes. The crystal structures of several proteins involved in Moco biosynthesis have been solved, and they mostly form oligomers⁴. This, together with biochemical evidence showing that many of these proteins are involved in the formation of a

heterogeneous complex, indicates that specific protein–protein interactions are crucial in the early stages of Moco biosynthesis⁵.

The new crystal structures³ are of Cnx1, which is homologous to gephyrin (in mammals) and MogA and MoeA (in bacteria). Once the molybdopterin moiety is formed at the end of the second stage, the metal ion has

to be prepared and inserted into it — and this is where Cnx1 comes into play. Cnx1 contains two domains, termed G and E, that catalyse the transfer and insertion of Mo into the molybdopterin⁶. Kuper *et al.*³ characterized domain G of Cnx1 structurally and biochemically under anaerobic conditions (because oxygen would cause some degradation). They report surprising results.

First, analysis of the structure of Cnx1 with a mutation in its active site revealed an adenosine moiety (AMP) covalently linked to molybdopterin. This chemical species has not been seen before, and the authors go on to show that it is Cnx1G itself that attaches the AMP to molybdopterin. This finding introduces a new step to Moco biosynthesis and is consistent with previous work⁵. Kuper

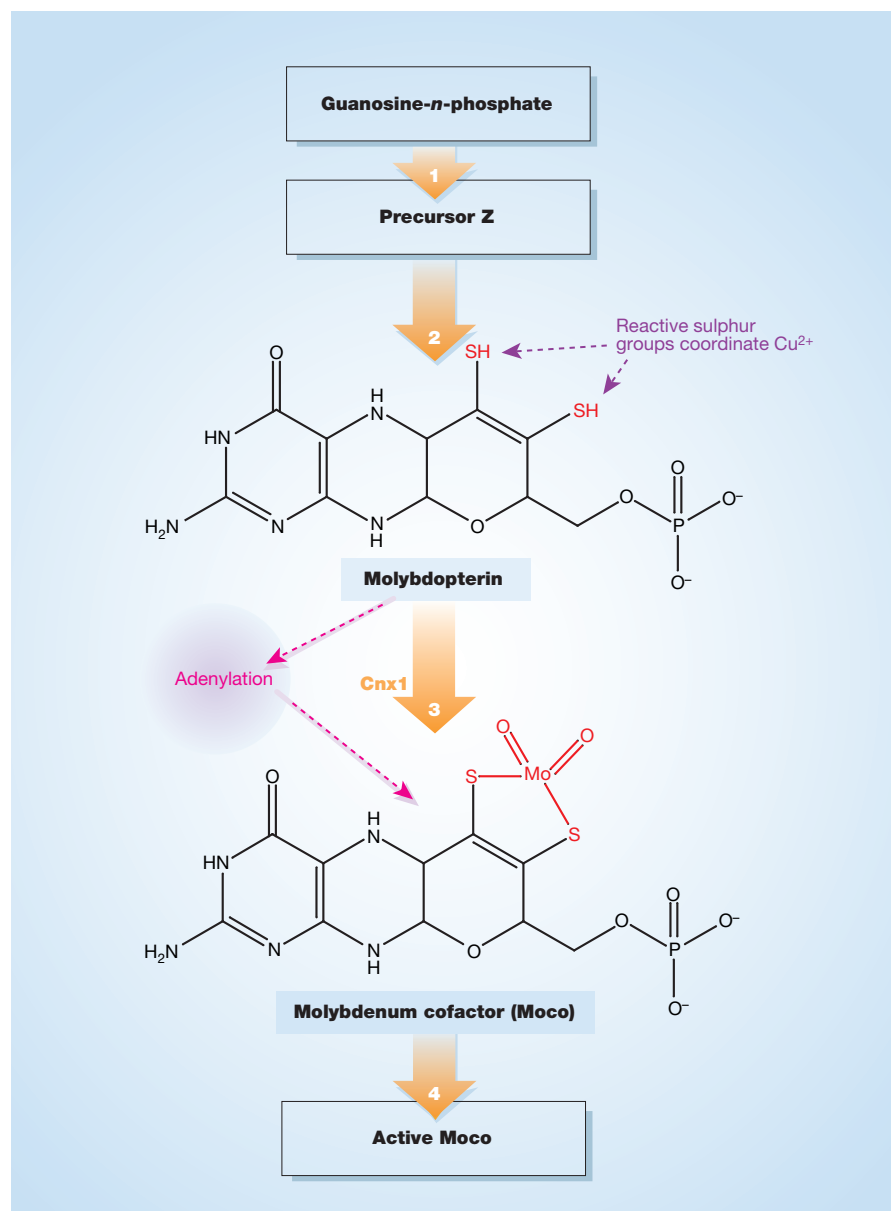


Figure 1 Four biosynthetic steps to the active molybdenum cofactor, Moco. In plants, the enzyme Cnx1 catalyses stage 3, where the molybdenum (Mo) ion is covalently bound to molybdopterin. Kuper *et al.*³ report two structures of the Mo-binding domain of Cnx1 that unexpectedly show the involvement of copper ions in the process, and reveal a new step in the pathway involving the addition of an adenosine moiety to molybdopterin.

*et al.*³ postulate that the product of Cnx1G, the molybdopterin-AMP, is cleaved by domain E of Cnx1, which also transfers the Mo to the molybdopterin sulphur groups to produce Moco. Understanding domain E and its interplay with domain G will now be crucial in completing the picture of this biosynthetic pathway.

Second — and perhaps most important — is the observation of copper binding to the two sulphur groups of the molybdopterin (Fig. 1). These sulphur atoms are potentially reactive and some mechanism is required to protect them until they actually coordinate Mo. The copper ion was not added to the protein during the experiment, but rather seems to have been picked up from the bacterial expression system used to produce the protein, which suggests that the levels of copper present in culture are sufficient for the purpose. The implication is that *in vivo* Cnx1 binds copper, probably for a combination of

reasons: a metal ion complexed to the reactive sulphur groups would not only protect them, but would also provide a mechanism to assist in the insertion of Mo into the cofactor as the metals are exchanged. The suggestion of a direct link between Cu and Mo biology is highly intriguing and will spur further investigations to delineate the relationship between these essential metal ions. ■

William N. Hunter is in the Division of Biological Chemistry and Molecular Microbiology, School of Life Sciences, University of Dundee, Dow Street, Dundee DD1 5EH, UK.
e-mail: w.n.hunter@dundee.ac.uk

1. Stiefel, E. I. in *Molybdenum Cofactors and Model Systems* (eds Stiefel, E. I., Coucouvanis, D. & Newton, N. W.) 1–19 (Am. Chem. Soc., Washington DC, 1993).
2. Kisker, C., Schindelin, H. & Rees, D. C. *Annu. Rev. Biochem.* **66**, 233–267 (1997).
3. Kuper, J. *et al. Nature* **430**, 803–806 (2004).
4. Rizzi, M. & Schindelin, H. *Curr. Opin. Struct. Biol.* **12**, 709–720 (2002).
5. Lake, M. W. *et al. Nature* **414**, 325–329 (2001).
6. Schwarz, G. *et al. J. Mol. Biol.* **312**, 405–418 (2001).

Climate change

Models change their tune

Thomas F. Stocker

Climate models are usually tuned to match observations. A new approach, in which the models are detuned instead, increases our confidence in projections of future warming.

In 2001, the Intergovernmental Panel on Climate Change published its report stating that the Earth will warm by between 1.4 °C and 5.8 °C by the end of the twenty-first century¹. Many argue that this range is too large to justify action to reduce the rising concentrations of greenhouse gases, principally carbon dioxide, in the atmosphere. Two factors are responsible for the wide estimated range: uncertainty in our understanding of the physical processes that influence climate and our ability to model them; and uncertainty over the so-called ‘emissions path’ — how the demographic, technological and ecological development of human societies will affect energy consumption and greenhouse-gas emissions.

Murphy and colleagues (page 768 of this issue)² have used a climate model to estimate the uncertainty associated with modelling the physical processes. Climate modellers use mathematical formulations, called parameterizations, to describe processes that cannot be resolved because of their scale or complexity, or that are not well understood. Murphy *et al.* quantify how the choice of parameterizations affects the results of modelling the evolution of temperature, precipitation and other climate variables when the greenhouse-gas concentration increases. The good news for the modelling community is that none of these choices alters the basic response of the model to a common

scenario, a doubling of CO₂ concentration: all simulations show consistent warming worldwide, most likely between 2 °C and 3 °C, with the characteristic amplification of warming at high latitudes of the Northern Hemisphere. So this is a robust result over a wide range of parameterizations.

Simplifications are unavoidable when trying to capture the climate system — with its cascades of scales in time and space — in a finite numerical model. Some of these parameterizations may be rigorously derived from fundamental physical principles³, but many contain poorly constrained parameters that are used to complete the mathematical construct. Although it is often true that (in John von Neumann’s words) “the justification of such a mathematical construct is solely and precisely that it is expected to work”, practical data from instruments must be used to set limits on the physically realistic range for these parameters.

Murphy and colleagues’ model² consists of a state-of-the-art atmospheric general circulation model coupled to a ‘slab’ ocean. The latter provides a rudimentary lower boundary condition for the atmospheric model, with lateral fluxes of heat being suggested and kept fixed, while sea surface temperatures are allowed to change. Ocean dynamics are not taken into account, which precludes projections of such considerations as changes in sea-ice cover in the Arctic and

Antarctic, or in the ocean uptake of CO₂. In predicting climate changes in the next several decades — the transient climate response — the ocean cannot be neglected^{4,5}. However, the equilibrium climate sensitivity, a key value in characterizing a climate model, can be reasonably estimated using this reduced set-up. Climate modellers are sometimes preoccupied with ‘tuning’ their models to achieve best agreement with observations, but Murphy *et al.* take the opposite tack: instead of aiming for a best estimate by careful tuning, they quantify the range of outcomes that results when parameter values are changed. This detuning is a neat way of perturbing some of the physics of the climate model.

Within given bounds, they varied 29 parameters, one by one, and analysed the results from 20-year simulations under present-day and doubled CO₂ conditions. In this linear approach, perturbing each parameter in turn, it is assumed that the effect of combinations of changed values of different parameters can be roughly estimated by adding their results. Murphy *et al.* define a climate prediction index, or quality index, which weighs the simulated fields against observations, and permits an objective determination of the critical parameters in the perturbation experiments. It is no surprise that fields associated with the hydrological cycle, particularly cloud processes, show the widest spread in this quality index. In consequence, the partitioning of the radiative fluxes (short-wave versus long-wave radiation, downwards versus upwards) varies widely when the parameterizations are perturbed. This was expected from earlier comparisons⁶; it indicates that the models must be improved if they are to be reconciled with high-precision data on components of the radiative balance⁷.

In a further step, Murphy *et al.* calculate probability density functions for climate sensitivity; this is defined as the estimated increase of global mean temperature in doubled CO₂ conditions. Earlier studies have produced such estimates with either small ensembles of comprehensive models⁴ or simplified climate models⁸. Depending on the various model simulations to which the quality index is applied, and giving more weight to those that show better agreement with observations, the warming in the 5–95% probability range is between 2.4 °C and 5.4 °C. In spite of the relatively wide range of systematic parameter variation, the fundamental response of the climate model is consistent with the range reported by the Intergovernmental Panel^{1,9}.

Uncertainty is inherent in any prediction. The public have become used to probabilities reported in weather forecasts. But the public, and decision makers, have yet to accept that a similar approach is necessary to guide steps to reduce and mitigate the impact

of climate change. Work such as that of Murphy *et al.* and other initiatives will reduce these uncertainties to a point where the dominating uncertainty remains the choice of emission path used in modelling. This is already the case for predictions that look at the second half of this century¹⁰. To shift that time horizon closer to the present will require improved climate models that can be compared in a systematic way¹¹, larger model ensembles for testing^{12,13}, and a systematic exploration of uncertainties in the radiative balance. ■

Thomas F. Stocker is in the Division of Climate and Environmental Physics, Physics Institute, University of Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland.
e-mail: stocker@climate.unibe.ch

Materials science

Flame-broiled alumina

Paul F. McMillan

A method for preparing aluminate glasses and glass-ceramic composites opens up new possibilities for generating mechanically strong structural components and high-hardness coatings.

Shattered glass and cracked dinnerware: ceramics are brittle materials that fail catastrophically when they are stressed. However, strategies are being developed for creating ceramics, including glasses, that are toughened against fracture¹. The best of these techniques use nanometre-size crystals that are formed by devitrification of a glassy matrix² — that is, the formation of ordered crystalline regions throughout the amorphous, non-crystalline glass. On page 761 of this issue, Rosenflanz *et al.*³ describe a new method: micrometre-size beads of aluminate glasses are formed using a flame-spray technique; the glassy beads are then sintered into bulk glasses; further heating produces toughened, hard ceramics in which nanocrystalline alumina-rich phases are dispersed throughout a glassy matrix.

Alumina — aluminium oxide, Al₂O₃ — is the basis of many important ceramic systems, such as optical fibres. Alumina ceramics are also the industry standard for structural applications at high temperatures, especially in chemically reactive environments. Fracture-resistant aluminas are ideal for use in gas turbines, as protective tiles on space probes, and as supports for catalytic reactor systems. These applications all involve forming the ceramics into complex, dense shapes. The best way to do that is to first prepare the materials in an amorphous form, such as glass, and then devitrify them.

However, aluminate glasses form only within a narrow range of compositions because of their high melting temperatures, low viscosity on melting, and the competing

1. Houghton, J. T. *et al.* (eds) *Climate Change 2001: The Science of Climate Change* (Cambridge Univ. Press, 2001).
2. Murphy, J. M. *et al.* *Nature* **430**, 768–772 (2004).
3. Saltzman, B. *Dynamical Paleoclimatology* (Academic, London, 2002).
4. Stott, P. A. & Kettleborough, J. A. *Nature* **416**, 723–726 (2002).
5. Raper, S. C. B., Gregory, J. M. & Stouffer, R. J. *J. Clim.* **15**, 124–130 (2002).
6. Wild, M., Ohmura, A., Gilgen, H., Morcrette, J. J. & Slingo, J. *J. Clim.* **14**, 3227–3239 (2001).
7. Philippona, R., Dürr, B., Marty, C., Ohmura, A. & Wild, M. *Geophys. Res. Lett.* **31**, L03202 (2004).
8. Knutti, R., Stocker, T. F., Joos, F. & Plattner, G.-K. *Nature* **416**, 719–723 (2002).
9. Cubasch, U. *et al.* in *Climate Change 2001: The Science of Climate Change* (eds Houghton, J. T. *et al.*) 525–582 (Cambridge Univ. Press, 2001).
10. Zwiers, F. W. *Nature* **416**, 690–691 (2002).
11. Covey, C. *et al.* *Glob. Planet. Change* **37**, 103–133 (2003).
12. Allen, M. R. & Stainforth, D. A. *Nature* **419**, 228 (2002).
13. www.climateprediction.net

crystallization kinetics of their components. No one has yet prepared a bulk glass from pure alumina liquid using standard melt-quenching techniques (the same is true of water and silicon). But Rosenflanz and colleagues' flame-spray method yields glassy microspheres containing more than 80% alumina³. The flame-spray technique consists of feeding oxide powder into a high-temperature hydrogen–oxygen flame and then cooling it rapidly (quenching) to produce the glassy beads for sintering. The authors suggest that the technique could be extended to produce glasses in other 'reluctant' glass-forming systems.

Aluminate glasses formed in combination with calcium or zirconium oxides, or rare-earth oxides (REOs), are remarkable materials: they are mechanically strong and hard, and have some of the highest sound speeds of any glassy system⁴. Interestingly, devitrified samples containing mechanically weak REO components have hardness values that are almost as high as those of dense polycrystalline alumina. This reflects the importance of the interfaces between the components for toughening the glass-ceramics^{3,5}: the energy of a crack propagating through the material may be deflected or absorbed at such an interface, adding to its strength. So controlling the overall texture in these materials is the key to designing and developing composite ceramics that are both hard and strong.

Rosenflanz *et al.*³ present another important observation from their study of the REO-containing composite gadolinium

aluminate (Al₂O₃–Gd₂O₃); this observation is linked to the physical properties of alumina-rich liquids, rather than to materials research. Molten aluminates are highly 'fragile', in that the relationship between their temperature and viscosity does not follow Arrhenius' law⁵. This is correlated with the large configurational entropy in these systems, reflecting the wide variety of structural states that they can adopt: for example, Al³⁺ ions occur in multiple coordination environments (tetrahedral, octahedral, 5-coordinate, and so on), and rare-earth metal ions have coordination numbers ranging from 6 to 10. The result is a rapidly changing structure within the high-temperature liquids.

An intriguing consequence of this behaviour is that aluminate liquids and glasses can exhibit 'polyamorphism' — different liquid phases or glasses are possible that, although they have the same chemical composition, have different structures and thermodynamic properties⁶. This results in an unusual new phenomenon for the liquid state: density- or entropy-driven liquid–liquid phase transitions can appear as the pressure and temperature change. These transitions are analogous to the temperature- and pressure-driven structural transformations seen in crystalline solids, such as the diamond–graphite transition or the transitions between the quartz, cristobalite, coesite and stishovite polyamorphs of silicon dioxide.

When heating their gadolinium-aluminate glass, Rosenflanz *et al.*³ noticed an exothermic — heat producing — feature that could indicate the presence of polyamorphism in the supercooled aluminate liquid. Similar features have already been observed among yttrium-aluminate (Y₂O₃–Al₂O₃) liquids that are known to exhibit low- and high-density polyamorphs possessing different values of hardness and other physical properties^{7,8}. Controlling the polyamorphism in glassy aluminates and the resulting ceramic microtextures provides a new way to tailor the mechanical properties, including hardness, of families of composite materials. ■

Paul F. McMillan is at the Davy–Faraday Laboratory, Royal Institution of Great Britain, and in the Department of Chemistry, Christopher Ingold Laboratory and Materials Chemistry Centre, University College London, London WC1H 0AJ, UK.
e-mail: p.f.mcmillan@ucl.ac.uk

1. Morgan, P. E. D. & Marshall, D. B. *Mater. Sci. Eng. A* **162**, 15–25 (1993).
2. Yip, S. *Nature* **391**, 532–533 (1998).
3. Rosenflanz, A. *et al.* *Nature* **430**, 761–764 (2004).
4. Yeganeh-Haeri, A. *et al.* *J. Non-Cryst. Solids* **241**, 200–203 (1998).
5. Weber, J. K. R., Felton, J. J., Cho, B. & Nordine, P. C. *Nature* **393**, 769–773 (1998).
6. McMillan, P. F. *J. Mater. Chem.* **14**, 1506–1512 (2004).
7. Weber, J. K. R. *et al.* *J. Am. Ceram. Soc.* **83**, 868–872 (2000).
8. McMillan, P. F., Wilson, M. & Wilding, M. C. *J. Phys. Cond. Matter* **15**, 6105–6121 (2003).

Structural biology

New role for Pauling's ribbons

Proc. Natl Acad. Sci. USA **101**, 11622–11627 (2004)

In 1951, Linus Pauling and Robert Corey proposed the existence of the α -helix and β -sheet — now known to be the two most common types of structural fold within proteins — by considering the stereochemistry of amino acids. They also proposed other protein conformations, which are less common, or even non-existent, *in vivo*. Roger S. Armen *et al.* suggest that one of the rarest of these is an intermediate in amyloid formation.

α -Helices can be left-handed or right-handed, but if amino acids alternate between these two orientations, the ribbons produced can form corrugated sheets, held together by hydrogen bonds. This α -sheet structure appeared in the authors' computer simulations of the acid-induced unfolding of four proteins, including lysozyme and prion protein, that are implicated in amyloid diseases — disorders in which proteins form insoluble 'amyloid' fibrils.

The protein regions that adopted α -sheet conformations are known to be important in fibril formation, suggesting a role for this ephemeral structure in amyloid assembly. Whether true or not, it shows that Pauling and Corey's model-building still illuminates research half a century on. **Christopher Surridge**

Particle physics

The Bs have it

Preprint at <http://arxiv.org/abs/hep-ex/0407057> (2004)

If physics were simple, matter and antimatter particles would mirror each other in every way. Their electrical charges are equal and opposite, they have identical masses, and they should decay at the same rate.

But in 1964, physicists found that a particle called a neutral *K*-meson decayed at a different rate from that of its antimatter counterpart. This 'charge parity (CP) violation' might explain why our Universe is built from an excess of matter rather than antimatter.

BaBar, an experiment at the Stanford Linear Accelerator, studies CP violation in particles called *B*-mesons, made by smashing together electrons and their antimatter counterparts positrons. In earlier studies at BaBar, and at a similar Japanese experiment called Belle, *B*-mesons were found to have a preference for existing in their matter rather than their antimatter form — an effect classed as indirect CP violation.

B. Aubert *et al.* at BaBar have now expanded on this result by counting the number of decay products formed from the *B*-mesons. They say that they found

910 examples of a *B*-meson decaying into a *K*-meson, but only 696 examples of the equivalent antiparticle process. This measurement is classed as direct CP violation, and is a much more subtle effect than previous experiments have tested for.

Mark Peplow

Conservation biology

Young blood

Conserv. Biol. **18**, 1078–1087 (2004)

A popular conservation strategy is the reintroduction of species into areas where they were once indigenous. Previous wisdom suggested that such projects should use adults, to get breeding under way as early as possible. But the authors of a fresh analysis propose that this dogma should be turned on its head.

Using a computer simulation, Alexandre Robert *et al.* calculated the overall fitness of the theoretical reintroduced populations seeded by either adults or juveniles. They found that the populations seeded by juveniles were less susceptible to the accumulation of deleterious mutations. To verify their model, the researchers then tested it against data from the successful real-life release of the griffon vulture (*Gyps fulvus fulvus*) into southern France, and found that their simulation mimicked the observed results.

The authors argue that future species-release projects should therefore use animals that are as young as possible. This will give time for selection to get to work before the animals begin breeding, resulting in a healthier population. The researchers acknowledge, however, that this is unlikely to be the case when using animals from captive breeding stock, as these populations are likely to be highly inbred and therefore unlikely to benefit from the purging effects of selection.

Michael Hopkin

Biomedical materials

Hold on to your teeth

Adv. Mater. **16**, 1071–1074 (2004)

Titanium implants are widely used for bone replacement: they are strong, corrosion-resistant and don't provoke an inflammatory response. But they don't work well at the body's surface, where they make contact with epithelial tissue, because epithelial cells don't stick securely to titanium. The weak adhesion leaves room for bacteria to infect the interface. This creates problems, for example, in the use of titanium for the roots of artificial teeth.

Masaki Uchida *et al.* have found a way to improve the adhesion between titanium and epithelial tissue. They imitate the biomineralization process that secures the mineral component of shell to its soft organic tissue: here, proteins that glue the two surfaces together are immobilized in the

mineral phase. Uchida *et al.* coat the surface of titanium with a composite of apatite — essentially, calcium phosphate, the mineral component of bone — and laminin, the adhesion protein responsible for binding real teeth to surrounding epithelial tissue. Apatite grows on titanium that is rendered reactive by treating it with sodium hydroxide and heat. When the mineral is grown from a laminin-containing solution, the protein becomes incorporated too, and human oral epithelial cells then grow readily on the composite's surface.

Philip Ball

Microbiology

Bacterial bridge

Proc. Natl Acad. Sci. USA **101**, 11448–11453 (2004)

Streptomyces are soil-dwelling bacteria that are best known for producing many naturally derived antibiotics. Unlike most bacteria, *Streptomyces* (pictured) have multicellular developmental stages. Bacterial colonies start growing in wet soil, and then grow away from the colony surface into the air, forming aerial hyphae. From these aerial structures, spores are formed, allowing the bacteria to disperse.



A key player in the formation of hyphae was identified more than a decade ago in *Streptomyces coelicolor*. This peptide, called SapB, acts as a surfactant and releases the surface tension at the air–water interface, allowing hyphae to grow into the air. Through a combination of genetic and structural analysis, Shinya Kodani *et al.* now propose that SapB is a lantibiotic-like peptide. Lantibiotics are antimicrobial peptides, and many are amphiphilic, with hydrophobic side chains on one side and hydrophilic side chains on the other. SapB is also amphiphilic, helping to explain its surfactant capability.

Surprisingly, the authors found that SapB does not seem to have antimicrobial activity, suggesting that it must have key structural differences from true lantibiotics. But the involvement of a lantibiotic-like peptide in a critical developmental stage of *Streptomyces* raises the possibility that lantibiotics evolved from this peptide.

Joanne Kotz

D. SCHARE/SPL

Corals' adaptive response to climate change

Shifting to new algal symbionts may safeguard devastated reefs from extinction.

The long-term response of coral reefs to climate change depends on the ability of reef-building coral symbioses to adapt or acclimatize to warmer temperatures, but there has been no direct evidence that such a response can occur. Here we show that corals containing unusual algal symbionts that are thermally tolerant and commonly associated with high-temperature environments are much more abundant on reefs that have been severely affected by recent climate change. This adaptive shift in symbiont communities indicates that these devastated reefs could be more resistant to future thermal stress, resulting in significantly longer extinction times for surviving corals than had been previously assumed.

Reef-building scleractinian (stony) corals obligately host dinoflagellate algal symbionts (genus *Symbiodinium*), whose loss during mass coral-reef 'bleaching' as a result of increased seawater temperatures is a significant threat to coral reef ecosystems worldwide¹. Although acclimatization and/or adaptation of reef coral hosts and/or their algal symbionts are recognized as potentially mitigating the frequency and severity of these bleaching events^{1–3}, no studies have been undertaken to test whether such responses actually occur on affected reefs.

The genus *Symbiodinium* is diverse, and many corals are relatively flexible in the type(s) of algal symbiont they contain, although one type is usually dominant in any given species and environment^{4–7}. Because corals containing different symbionts can vary in their sensitivity to bleaching^{4,8} and can modify their symbiont communities in response to environmental change^{4,7,9,10}, we investigated whether severe bleaching and mortality can select for stable host-symbiont combinations that are more thermally tolerant, raising the overall bleaching resistance of the reef as a result.

We undertook molecular surveys of *Symbiodinium* in shallow (less than 7 m depth) scleractinian corals from five locations in the Indo-Pacific that had been differently affected by the 1997–98 El Niño–Southern Oscillation (ENSO) bleaching event (Fig. 1; for sample details, see supplementary information). For these large-scale comparisons, restriction-fragment length polymorphisms in the symbionts' large-subunit ribosomal DNA were used to distinguish *Symbiodinium* in clades A, C or D^{5,9}.

In Panama, we surveyed ecologically dominant corals in the genus *Pocillopora* before, during and after ENSO bleaching. Colonies containing *Symbiodinium* in clade D were already common (43%) in 1995 and

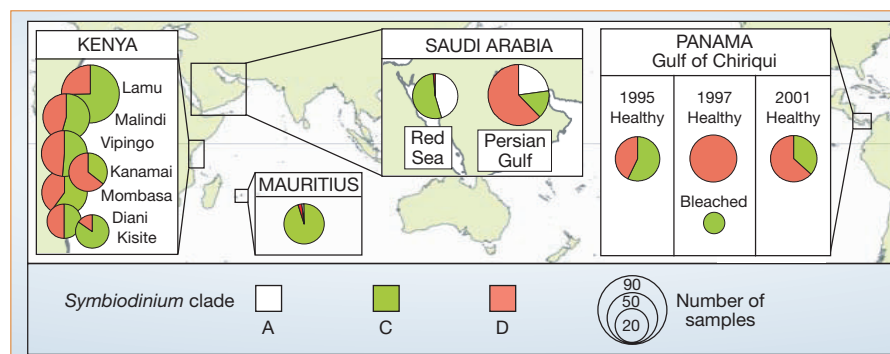


Figure 1 Distribution of *Symbiodinium* algae in shallow-water (less than 7 m depth) scleractinian corals from Kenya, Mauritius, Saudi Arabia and Pacific Panama. Pie charts show distribution of symbionts by site, except those for Panama, which show the distribution of symbionts at a single site before (1995: all colonies healthy), during (1997: some colonies healthy, others severely bleached) and after (2001: all colonies healthy) the 1997–98 El Niño. Size of pie charts is scaled to the square-root of sample size to reflect an equal area for each sample, as indicated by the inset scale. See supplementary information for sampling details for each country.

were unaffected by bleaching in 1997, while colonies containing clade C bleached severely⁸. By 2001, colonies containing clade D had become dominant (63%) on these reefs.

We also surveyed corals from the Persian (Arabian) Gulf that experience extremely high seasonal temperatures (> 33 °C). Even these reefs were severely bleached in 1998 (ref. 11), when temperatures in some locations exceeded 38 °C. We compared these with corals from the Red Sea that do not experience such seasonal temperature highs (typically 29 °C) and which were relatively unaffected in 1998 (ref. 11). We found that, in 2000–01, Gulf reefs were dominated by the same clade-D symbionts as were found in Panama (62% of scleractinian colonies), whereas Red Sea reefs only rarely contained these symbionts (1.5% of colonies).

This comparison was repeated in the western Indian Ocean in 2000–02: corals from reefs in Kenya that were severely bleached by the combined effects of ENSO and the Indian Ocean Dipole in 1998 were compared with corals from Mauritius that had fortuitously escaped significant bleaching during the same event¹¹. The findings were similar: in Kenya, 15–65% of colonies contained clade D, depending on site, compared with only 3% of colonies in Mauritius.

Taken together, these results indicate that corals containing thermally tolerant *Symbiodinium* in clade D are more abundant on reefs after episodes of severe bleaching and mortality, and that surviving coral symbioses on these reefs more closely resemble those found in high-temperature environments.

We have yet to determine the processes by which these symbiont shifts occur in coral populations (for example, there could be differential mortality of coral hosts and/or bleaching-induced symbiont change

in individual colonies¹²). It is possible that affected reefs may revert to their original symbiont communities over time if they do not experience repeat warming episodes¹³. We also recognize that some corals (most notably the genus *Porites*) remain common on devastated reefs, even though they do not often contain clade-D symbionts. Nevertheless, we propose that the symbiont changes described here are a common feature of severe bleaching and mortality events, and predict that these adaptive shifts will increase the resistance of these recovering reefs to future bleaching.

Andrew C. Baker*†, Craig J. Starger‡, Tim R. McClanahan*, Peter W. Glynn§

*Marine Program, Wildlife Conservation Society, Bronx, New York 10460, USA
e-mail: abaker@wcs.org

†Center for Environmental Research and Conservation and ‡Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, New York 10027, USA
§Rosenstiel School of Marine and Atmospheric Science, University of Miami, Florida 33149, USA

1. Hughes, T. P. *et al.* *Science* **301**, 929–933 (2003).
2. Coles, S. L. & Brown, B. E. *Adv. Mar. Biol.* **46**, 183–223 (2003).
3. Sheppard, C. R. C. *Nature* **425**, 294–297 (2003).
4. Rowan, R., Knowlton, N., Baker, A. C. & Jara, J. *Nature* **388**, 265–269 (1997).
5. Baker, A. C. *Annu. Rev. Ecol. Evol. Syst.* **34**, 661–689 (2003).
6. Lajeunesse, T. C. *Mar. Biol.* **141**, 387–400 (2002).
7. Little, A. F., Van Oppen, M. J. H. & Willis, B. L. *Science* **304**, 1492–1494 (2004).
8. Glynn, P. W., Maté, J. L., Baker, A. C. & Calderón, M. O. *Bull. Mar. Sci.* **69**, 79–109 (2001).
9. Baker, A. C. *Nature* **411**, 765–766 (2001).
10. Toller, W. W., Rowan, R. & Knowlton, N. *Biol. Bull.* **201**, 360–373 (2001).
11. Wilkinson, C. R. (ed.) *Status of Coral Reefs of the World: 2002* (Australian Institute of Marine Science, Queensland, 2002).
12. Buddemeier, R. W. & Fautin, D. G. *Bioscience* **43**, 320–326 (1993).
13. Ware, J. R., Fautin, D. G. & Buddemeier, R. W. *Ecol. Model.* **84**, 199–214 (1996).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Coral bleaching

Thermal adaptation in reef coral symbionts

Many corals bleach as a result of increased seawater temperature, which causes them to lose their vital symbiotic algae (*Symbiodinium* spp.) — unless these symbioses are able to adapt to global warming, bleaching threatens coral reefs worldwide^{1–3}. Here I show that some corals have adapted to higher temperatures, at least in part, by hosting specifically adapted *Symbiodinium*. If other coral species can host these or similar *Symbiodinium* taxa, they might adapt to warmer habitats relatively easily.

Around Guam, species of the coral genus *Pocillopora* each associate with at least two *Symbiodinium* taxa, one of which, according to ecological data⁴, seems to be more tolerant of high temperature. I tested whether this could be the case by comparing photosynthetic responses of the taxa, labelled according to their genotype, *Symbiodinium C* and *Symbiodinium D* (ref. 4) (for methods, see supplementary information). I measured the maximum quantum yield of photosystem II (PSII) as the ratio of variable chlorophyll fluorescence (F_v/F_m)⁵ in *P. verrucosa*. In *P. damicornis*, I measured photosynthesis from oxygen flux.

Symbiodinium C and *D* respond in opposite ways to temperature, as indicated by their differing F_v/F_m (Fig. 1a). Compared with a control temperature of 28.5 °C, a temperature of 31.3 °C did not affect *Symbiodinium C*, but it increased F_v/F_m in *Symbiodinium D*; a temperature of 32.0 °C decreased F_v/F_m in *Symbiodinium C*, whereas *Symbiodinium D* maintained an increased F_v/F_m . Although F_v/F_m was similar in *Symbiodinium C* and *D* at 28.5 °C, at 32.0 °C *Symbiodinium C* could be identified by its lower F_v/F_m . After the temperature treatments, corals were kept at 28.5 °C; after three and four days, F_v/F_m in treated *Symbiodinium C* remained lower than in controls ($P=0.02$) and unchanged from the value recorded at 32.0 °C ($P>0.2$), whereas F_v/F_m in control and treated *Symbiodinium D* had become similar ($P>0.2$; Wilcoxon paired-sample tests).

A long-lasting decrease in F_v/F_m , as observed in *Symbiodinium C*, indicates that chronic photoinhibition resulted from damage to PSII (refs 5–7). Repeated measures (Wilcoxon paired-sample tests) show that the decrease in F_v/F_m in *Symbiodinium C* at 32.0 °C compared with 31.3 °C ($P=0.02$) was accompanied by a 20% increase ($P=0.02$) in the minimum chlorophyll fluorescence in the dark-acclimated state (F_o) and no change ($P>0.5$) in F_m , confirming chronic photoinhibition⁵. Over the same time, both F_o and F_m decreased by 13% ($P=0.02$) in control *Symbiodinium C*, suggesting an increase in photoprotection⁷. F_o and F_m did

not change in *Symbiodinium D* under control or treatment conditions ($P\geq 0.2$).

Whereas chronic photoinhibition of *Symbiodinium C* indicates temperature sensitivity and predicts coral bleaching^{1,6,7}, the increased F_v/F_m in treated *Symbiodinium D* indicates photoprotection. For *Symbiodinium D*, the relationship between F_v/F_m and irradiance

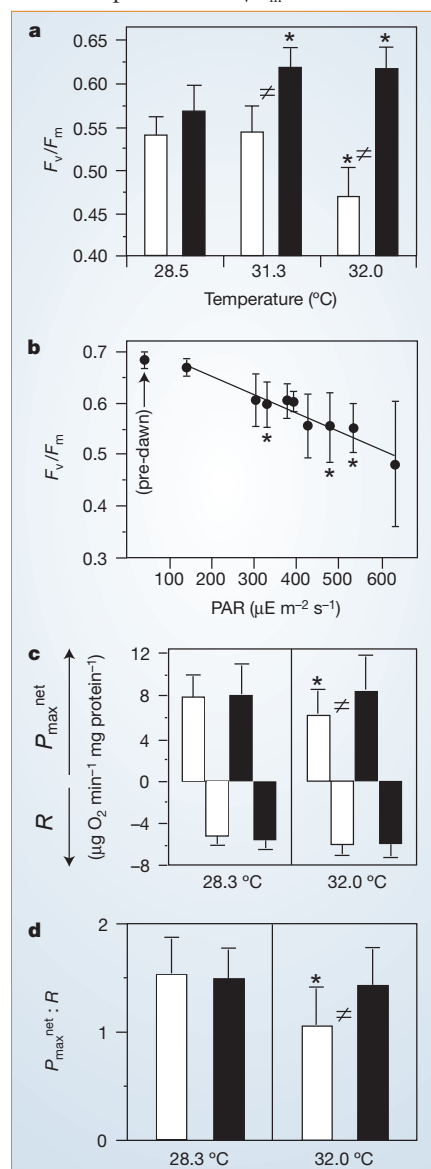


Figure 1 Photosynthesis in corals hosting *Symbiodinium C* (white bars) or *D* (black bars). **a**, Afternoon values of maximum quantum yield of photosystem II (F_v/F_m) in *Pocillopora verrucosa* (mean and s.d.; $n=7$); asterisks indicate differences ($P < 0.05$) between corals at 28.5 °C and the same corals at 31.3 °C or 32.0 °C; inequalities indicate differences ($P < 0.05$) between *Symbiodinium C* and *D* (both, Wilcoxon paired-sample tests). **b**, Afternoon values of F_v/F_m (mean $\pm$ s.d., $n=7$) in *P. verrucosa* hosting *Symbiodinium D* at 28.5 °C, plotted against daily average irradiance (PAR, photosynthetically available radiation; between 10:00 and 14:00); slope is not equal to zero (t -test, $P < 0.001$); pre-dawn (three-day average) shows overnight recovery of F_v/F_m compared with preceding afternoons (asterisks: $P=0.02$, Wilcoxon paired-sample test). **c**, P_{max}^{net} (positive) and R (negative, measured in the dark) of *P. damicornis* (mean and s.d.; $n=9$). **d**, $P_{max}^{net}:R$ from data summarized in **c**. In **c**, **d**, asterisks as in **a**; inequalities indicate differences between *Symbiodinium C* and *D* ($P < 0.05$, Mann–Whitney U -test).

exposure, which quantifies dynamic photoinhibition (reversible and protective) of PSII (refs 8, 9; Fig. 1b), shows that increased temperature mimicked a 30% decrease in habitat irradiance at 28.5 °C. Photoprotection by increased temperature reflects the temperature dependence of photosynthetic pathways¹⁰. Thus, I conclude that *Symbiodinium D* is a high-temperature specialist. Plant models⁹ indicate that photoinhibition similar to that relieved by warmer temperatures in *Symbiodinium D* reduces daily carbon gain by 6–10%.

Oxygen-flux measurements independently support these conclusions and extend them to another host species at the whole-coral level. Increased temperature affected only corals hosting *Symbiodinium C*: maximum net photosynthesis (P_{max}^{net}) decreased; respiration (R) was not affected (Fig. 1c). At the higher temperature, the ratio of P_{max}^{net} to R ($P_{max}^{net}:R$) decreased by 31%, making corals hosting *Symbiodinium C* less autotrophic than corals hosting *Symbiodinium D* (Fig. 1d). Temperature did not affect numbers or the chlorophyll *a* of *Symbiodinium C* or *D* ($P\geq 0.5$, Wilcoxon paired-sample tests), so the decreased autotrophy did not result from lost symbionts.

Symbiodinium can differ physiologically owing to their acclimatization to different environments³, which probably include different host species. However, because I controlled for these variables, the differences observed here are regarded as intrinsic symbiont adaptations that apparently contribute significantly to whole-coral physiology. Adaptation to higher temperature in *Symbiodinium D* can explain why *Pocillopora* spp. hosting them resist warm-water bleaching whereas corals hosting *Symbiodinium C* do not (personal observations). It can also explain why *Pocillopora* spp. living in frequently warm (more than 31.5 °C) habitats host only *Symbiodinium D* (ref. 4), and, perhaps, why those living in cooler habitats predominantly host *Symbiodinium C* (ref. 4). These observations, which may apply to other corals¹¹, indicate that symbiosis recombination¹² may be one mechanism by which corals adapt, in part, to global warming³.

Rob Rowan

University of Guam Marine Laboratory, Mangilao, Guam 96923, USA

e-mail: rrowan@uog9.uog.edu

- Hoegh-Guldberg, O. *Mar. Freshwat. Res.* **50**, 839–866 (1999).
- Hughes, T. P. *et al.* *Science* **301**, 929–933 (2003).
- Coles, S. L. & Brown, B. E. *Adv. Mar. Biol.* **46**, 183–223 (2003).
- Rowan, R. *Ecology* (submitted).
- Maxwell, K. & Johnson, G. N. *J. Exp. Bot.* **51**, 659–668 (2000).
- Warner, M. E., Fitt, W. K. & Schmidt, G. W. *Proc. Natl Acad. Sci. USA* **96**, 8007–8012 (1999).
- Jones, R. J. & Hoegh-Guldberg, O. *Plant Cell Envir.* **24**, 89–99 (2001).
- Long, S. P., Humphries, S. & Falkowski, P. G. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **45**, 633–662 (1994).
- Werner, C. *et al.* *Plant Cell Envir.* **24**, 27–40 (2001).
- Huner, N. P. A., Oquist, G. & Sarham, F. *Trends Plant Sci.* **3**, 224–230 (1998).
- Baker, A. C. *Annu. Rev. Ecol. Evol. Syst.* **34**, 661–689 (2003).
- Rowan, R. & Powers, D. A. *Science* **251**, 1348–1351 (1991).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Genetic analysis of genome-wide variation in human gene expression

Michael Morley^{1,3*}, Cliona M. Molony^{2*}, Teresa M. Weber^{1,3}, James L. Devlin², Kathryn G. Ewens², Richard S. Spielman² & Vivian G. Cheung^{1,2,3}

¹Department of Pediatrics and ²Department of Genetics, University of Pennsylvania,

³The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, USA

* These authors contributed equally to this work

Natural variation in gene expression is extensive in humans and other organisms, and variation in the baseline expression level of many genes has a heritable component. To localize the genetic determinants of these quantitative traits (expression phenotypes) in humans, we used microarrays to measure gene expression levels and performed genome-wide linkage analysis for expression levels of 3,554 genes in 14 large families. For approximately 1,000 expression phenotypes, there was significant evidence of linkage to specific chromosomal regions. Both *cis*- and *trans*-acting loci regulate variation in the expression levels of genes, although most act *in trans*. Many gene expression phenotypes are influenced by several genetic determinants. Furthermore, we found hotspots of transcriptional regulation where significant evidence of linkage for several expression phenotypes (up to 31) coincides, and expression levels of many genes that share the same regulatory region are significantly correlated. The combination of microarray techniques for phenotyping and linkage analysis for quantitative traits allows the genetic mapping of determinants that contribute to variation in human gene expression.

The expression level of many genes shows abundant natural variation in species from yeast to humans^{1–6}. This trait, the 'gene expression phenotype'⁷, also shows familial aggregation^{5,6,8} and simple segregation patterns in yeast², suggesting an inherited contribution. Here, we extend our analysis⁶ by genetic mapping of regulatory elements that influence the baseline level of gene expression in human cells. Our goal is to identify determinants whose 'targets' are the genes with regulated expression.

We used microarrays to measure the baseline expression levels of genes in immortalized B cells from members of Centre d'Etude du Polymorphisme Humain (CEPH) Utah pedigrees⁹. For each of the ~8,500 genes on the array, we estimated the variance of expression level among unrelated individuals (94 CEPH grandparents) and the mean of variance of array replicates (two array replicates per individual). We restricted our analysis to genes with greater expression variation between individuals than between replicates (within individuals); these 3,554 most variable expression phenotypes are the quantitative traits that we mapped to chromosomal locations by genome scans.

Genotypes for genetic markers (single nucleotide polymorphisms; SNPs) were obtained from The SNP Consortium¹⁰. We used the computer program S.A.G.E. v. 4.5 (ref. 11) to carry out genome-wide linkage analysis for the 3,554 expression phenotypes in 14 CEPH families. The analysis gives the strength of the evidence for linkage at each map position in the form of a *t*-value¹², with associated point-wise significance level.

We selected expression phenotypes for further analysis using two different levels of stringency for evidence of linkage (that is, for a regulator of expression phenotype). For the more stringent level, we used a threshold of *t* > 5 from the S.A.G.E. analysis; in our sample of families, this corresponds to a point-wise *P*-value of <4.3 × 10^{−7} (a logarithm of odds (lod) score of ~5.3). For such a finding, the corresponding genome-wide significance level¹³ (see Methods) is approximately 0.001. Applying this genome-wide threshold to 3,554 scans we would expect only 3.5 genome scans to show any linkage evidence with a *P*-value this extreme by chance. Instead we found 142 expression phenotypes with evidence for linkage beyond the *P*-value threshold, and in some cases far beyond, so we conclude that false-positive linkage findings are at most a small fraction of the significant results. The expression phenotypes with the most

significant evidence of linkage are shown in Table 1.

In order to include additional phenotypes, we relaxed the stringency in some of the analyses by lowering the threshold to *t* > 4, which corresponds to a point-wise *P*-value of <3.7 × 10^{−5} (lod ~3.4) and approximately *P* = 0.05 genome-wide. There are 984 expression phenotypes that exceed this threshold, far more than the ~178 false positives expected by chance.

Cis and *trans* regulators of expression phenotypes

We consider the regions that are linked to the expression levels to be regulatory regions or 'regulators' of the expression phenotypes (of the target genes). We examined the regulatory regions for the 142 expression phenotypes with the most significant evidence of linkage, and for each quantitative phenotype we distinguished between apparently *cis*- and *trans*-acting regulators. We restricted the category of *cis* regulators to those that mapped within 5 megabases (Mb) of the target gene. This relatively large region was chosen to allow for imprecision of linkage and to include long-range regulators, as some *cis*-acting regulators act over megabase distances^{14,15}. All other significant linkage represents *trans* regulators. By these definitions of *cis* and *trans*, we found the following distribution of phenotypes: 27 (19%) have only a *cis*-acting transcriptional regulator, 110 (77.5%) have only a *trans*-acting regulator, and 5 (3.5%)

Table 1 Expression phenotypes with the strongest evidence of linkage from genome scans

<i>P</i> -value	Gene	Location	<i>Cis/trans</i>
<10 ^{−11}	ICAP-1A	2q25	<i>Cis</i> *
<10 ^{−11}	TM7SF3	12p11	<i>Cis</i> *
<10 ^{−10}	HSD17B12	11p11	<i>Cis</i>
<10 ^{−10}	CHI3L2	1p12	<i>Cis</i>
<10 ^{−10}	PSPHL	7p11	<i>Cis</i>
<10 ^{−10}	DSCR2	21q22.2	<i>Trans</i>
<10 ^{−10}	CBR1	21q22.1	<i>Trans</i>
<10 ^{−10}	HOMER1	5q14	<i>Trans</i>
<10 ^{−9}	DDX17	22q13	<i>Cis</i>
<10 ^{−9}	ZP3	7q11	<i>Cis</i>
<10 ^{−9}	IL16	15q25	<i>Cis</i>
<10 ^{−9}	ALG6	1p31	<i>Trans</i>
<10 ^{−9}	TNFRSF11A	18q22	<i>Trans</i>

*The most extreme *P*-values occurred at SNPs located >5 Mb from the gene, but linkage evidence within 5 Mb of target gene was also highly significant (*P* < 4.3 × 10^{−7}).

have two regulators (two phenotypes with a *cis*- and a *trans*-acting regulator, and three phenotypes with two *trans*-acting regulators).

Examples of the genome scan results for several expression phenotypes are shown in Fig. 1. We detected multiple regulators for only a small proportion of expression phenotypes, possibly because individual regulators make a smaller contribution when several regulators influence the expression level of a gene. Thus, some true regulators would not meet our criterion of $P < 4.3 \times 10^{-7}$ for evidence of linkage. We therefore also examined the 984 expression phenotypes with at least one marker significant at the reduced stringency of $P < 3.7 \times 10^{-5}$. Among these, we found 164 (16%) with multiple regulators of expression level, an appreciably higher percentage than the 3.5% found with the more stringent threshold. Multiple *trans*-acting regulators were found for 152 phenotypes, with both *cis*- and *trans*-acting regulators for the remaining 12.

Master regulators of transcription

In addition to genomic regions with regulators that affect single phenotypes *in cis* or *in trans*, we found genomic regions containing transcriptional regulators that influence multiple expression phenotypes. We divided the autosomal genome into 491 windows of 5 Mb and determined the number of regulators mapping to each window. We began by examining the regulators for the 142 expression phenotypes with $P < 4.3 \times 10^{-7}$. We found windows that contained many more 'hits' than expected by chance. If regulators for these phenotypes were distributed at random across the genome, the probability of six or more hits per window would be

less than 6×10^{-5} and we would not expect to see any windows with more than four hits. Instead, we found two hotspots with six or more hits ($P < 0.03$ after Bonferroni correction): seven phenotypes mapped to one window on chromosome 14, and six phenotypes mapped to one window on chromosome 20.

When we relaxed our linkage criterion to include the 984 regions with $P < 3.7 \times 10^{-5}$, we found many more expression phenotypes whose regulation mapped to shared hotspots. The two regions indicated above contain *trans*-acting regulators for the most expression phenotypes (Fig. 2a). Regulation for 31 of the 984 expression phenotypes mapped to the 5-Mb window on chromosome 14 (14q32), and regulation for 25 phenotypes mapped to the window on chromosome 20 (20q13).

We consider the existence of hotspots to be evidence for master regulators of the baseline level of gene expression. The mapping was done without considering possible relationships among phenotypes, but the shared expression control regions suggest co-regulation. We therefore examined the correlation in expression levels of the 31 and 25 target genes corresponding to the two master regulatory regions. The expression levels in 94 CEPH grandparents were used. In permutation tests with 1,000 replications, we found that the pairwise correlation between any two expression phenotypes did not exceed 0.52. We therefore set 0.52 as the threshold for correlation by chance (nominal $P = 0.001$). Hierarchical clustering was used to summarize these results graphically and group genes by the correlations of their expression levels. We looked for clusters of expression phenotypes whose members have pairwise correlations that all exceed 0.52. Among the 31 target genes whose regulators

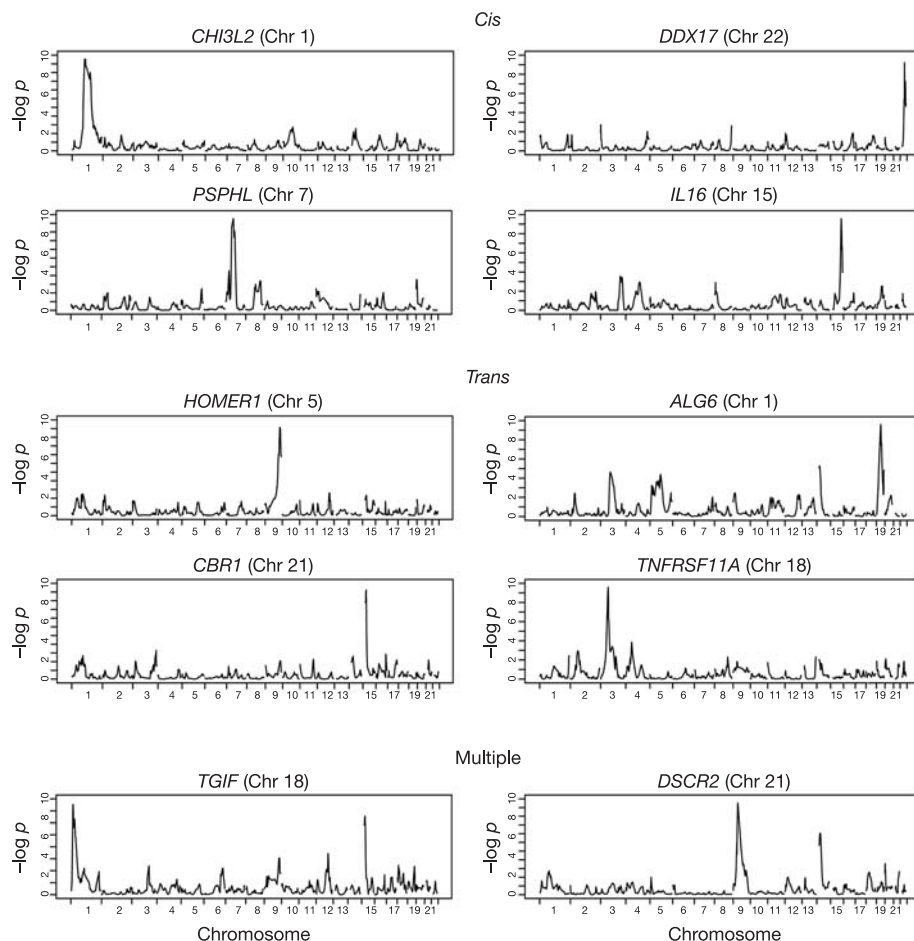


Figure 1 Genome scans for ten expression phenotypes. Chromosomal location of the regulated (target) gene is given in parentheses. The top eight panels show examples of

linkage with *cis*- or *trans*-acting transcriptional regulators. The bottom two panels show examples of phenotypes regulated by several unlinked genetic determinants.

mapped to the chromosome 14 hotspot, we found one such regulated cluster with 14 genes, and three additional clusters each with two genes (Fig. 2b). Similarly, among the 25 phenotypes whose regulators mapped to the chromosome 20 region, we found one cluster of four and two clusters of two genes whose members have pairwise correlations that all exceed 0.52. The correlation in expression level of these genes supports the observation that they share common transcriptional regulators. However, the regulatory regions defined by mapping are still large, and there might be subgroups of co-regulated phenotypes that are influenced by distinct, but very closely linked, regulators.

Some sets of closely linked genes are influenced by the same *cis* regulators, and have correlated expression profiles^{16–18}. In our data,

some target genes whose expression levels map to a *trans*-acting master regulatory region are also located very close to each other. For example, among the target genes whose expression maps to the regulatory region on chromosome 14 are four genes (*MMP24*, *C20orf24*, *RPN2* and *TOP1*) found in a 6-Mb region of chromosome 20 (UCSC Genome Browser, version hg15). In addition, among the target genes of the regulatory region on chromosome 20 are two genes (*ITM2B*, *RBI*) separated by less than 50 kilobases (kb) on chromosome 13. These observations reflect a complex regulatory network where master transcriptional regulators affect baseline expression levels of many genes that have similar expression profiles, and in some cases, reside close together on human chromosomes.

Family and population association analysis

Unlike linkage *in trans*, *cis* linkage of phenotypes immediately suggests a small region containing the regulatory element. This expectation led us to carry out follow-up studies with markers at several of the regulated genes. Among the 27 phenotypes with *cis* regulators (at $P < 4.3 \times 10^{-7}$), 17 were followed up by typing two or more additional SNP markers within or near the target gene. In each case, the expression data were used for both family-based and population-based analysis (Table 2). Analysis of the members of the 14 CEPH pedigrees by the Quantitative Transmission Disequilibrium Test (QTDT)¹⁹ showed significant evidence ($P < 0.01$) for the combined presence of linkage and association at 14 (82%) of these 17 genes, strengthening our conclusions in several ways. First, the QTDT results confirm the mapping in these cases to the target genes. Second, the results therefore support the inferred regulation *in cis*. Finally, the results also imply differential allelic expression (see below).

These conclusions were extended by regression analysis of data from 94 unrelated CEPH grandparents. Marked associations were found for many genes between expression phenotype and closely linked SNPs. Figure 3 shows examples with leukocyte-derived arginine aminopeptidase (LOC64167 or *LRAP*) and 3-ketoacyl-CoA reductase (*HSD17B12*). The corresponding results for linear regression analysis are given in Table 2. Among the 17 phenotypes tested, the same 14 found to be significant by QTDT showed highly significant evidence ($P < 0.005$) for population association between gene expression level and a SNP located within or near the gene (Table 2), directly demonstrating differential allelic expression. For two genes (*TM7SF3*, *ICAP-1A*) that did not show significant association, the linkage peaks were exceptionally broad as well as high, and spanned more than 10 Mb. Although evidence at the target gene itself was also statistically significant, the highest linkage peak was located >5 Mb away.

Differential allelic expression

The degree of differential allelic expression detected varies considerably. The largest effect was found for the phosphoserine phosphatase-like (*PSPHL*) gene, where there was an approximately eightfold difference in mean expression level between individuals homozygous for different alleles of a SNP marker (rs6700). In contrast, for a SNP (rs7176604) in the coding region of cathepsin H (*CTSH*), the fold difference between CC and TT homozygotes was 1.44 (Table 2). To follow up this latter finding, we used allele-specific quantitative real-time polymerase chain reaction (qRT-PCR) to compare the expression of the two alleles of that marker in 30 heterozygous individuals. We found similar allelic differences in expression level (mean fold difference = 1.6; s.d. = 0.45). The QTDT results for *CTSH* strongly support this finding ($P = 3 \times 10^{-6}$). Thus, several related approaches confirm the allelic differences *in cis*, and imply linkage disequilibrium between a SNP genotype and a nearby determinant that influences expression phenotype.

The region of linkage disequilibrium associated with differential allelic expression of some phenotypes extends for large distances, up

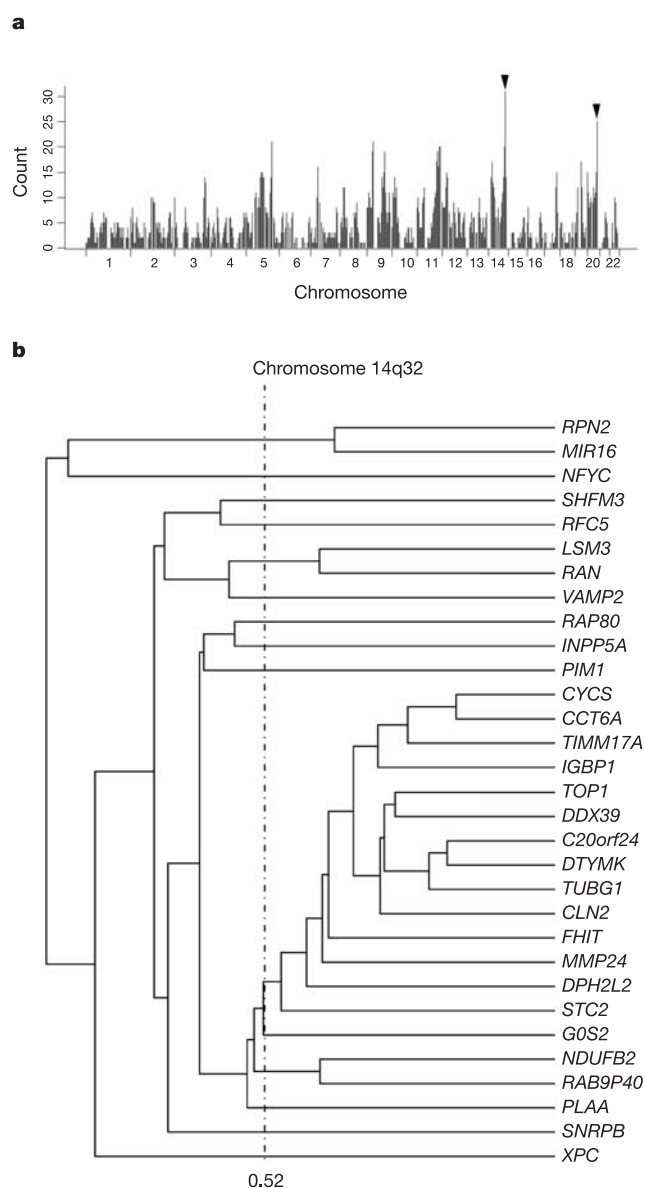


Figure 2 Master transcriptional regulators. **a**, Distribution of significant linkage peaks ($P < 3.7 \times 10^{-5}$) in 5-Mb windows across the autosomal genome. Arrowheads indicate the two windows (located on chromosomes 14 and 20) that contain regulators for the most expression phenotypes. **b**, Dendrogram representing hierarchical clustering of genes whose transcriptional regulators map to one 5-Mb window. Target genes for the hotspot of regulation on chromosome 14q32 are shown. Expression levels of genes with branches connected to the right of the dotted line are correlated at $P < 0.001$ (see text).

to several hundred kilobases. For example, two SNPs at *HSD17B12*, separated by ~172 kb, show nearly identical correlations with gene expression level (Fig. 3). This strong linkage disequilibrium makes it difficult to narrow down the region that contains the sequence variant(s) responsible for variation in the expression level of *HSD17B12*. In order to determine whether studying other populations (which may have different linkage disequilibrium structure) can solve this problem, we examined the linkage disequilibrium pattern for the same genomic region in African-Americans and found much less linkage disequilibrium for the SNPs shown. Standardized linkage disequilibrium coefficient D' is estimated as 1.0 in CEPH, but only 0.116 in African-Americans. In general, smaller linkage disequilibrium will make it easier to localize determinants²⁰.

For the phenotypes listed in Table 2, we estimated how much of the variation in expression phenotype could be attributed to *cis*-acting regulators. We used the results of the linear regression analysis, and calculated R^2 (the customary estimate of variation

explained by regression; see Methods). The last column of Table 2 shows the estimates obtained. These may be thought of as the 'heritability' attributable to the chromosomal region that is in linkage disequilibrium with the SNP tested, and therefore indicate what part of the variation in expression is influenced by *cis*-acting genetic determinants. For four of the genes in Table 2, this fraction is large—greater than 0.50. On the other hand, the fraction not explained in this way is also of interest, as it includes non-genetic causes (environmental differences, measurement variation) and possibly other genetic differences not in linkage disequilibrium with the SNP.

Discussion

Our study combined microarray expression data with publicly available SNP genotype data, and applied genome-wide mapping techniques to identify the chromosomal regions linked to the gene expression phenotypes. Level of gene expression is thus a trait like many others, and is amenable to genetic analysis. The classical

Table 2 Properties of genes whose expression level is regulated by *cis*-acting determinants

Gene	Location	Peak (<i>cis</i>) P -value (genome scan)	SNP (rs or ABI hCV identifier)	Highest/lowest ratio†	Population association‡ P -value	QTD P -value	Variation explained (R^2)
LOC64167	5q15	1×10^{-7}	4869311	7.02	2.3×10^{-19}	6×10^{-24}	0.60
<i>HSD17B12</i>	11p11	2×10^{-11}	1061810	1.68	5.9×10^{-18}	6×10^{-22}	0.57
<i>RPS26</i>	12q13	2×10^{-9}	1506440	1.47	1.7×10^{-17}	1×10^{-15}	0.55
<i>IRF5</i>	7q32	2×10^{-8}	2280714	2.33	2.0×10^{-16}	2×10^{-19}	0.52
<i>CSTB</i>	21q22	2×10^{-9}	26539999	1.74	2.9×10^{-12}	1×10^{-13}	0.42
<i>PSPHL</i>	7p11	3×10^{-11}	6700	8.43	3.5×10^{-12}	3×10^{-14}	0.41
<i>CHI3L2</i>	1p12	3×10^{-11}	755467	3.69	1.7×10^{-11}	8×10^{-12}	0.39
<i>CPNE1</i>	20q11	1×10^{-7}	6060516	2.57	6.8×10^{-11}	4×10^{-13}	0.38
<i>CTBP1</i>	4p16	2×10^{-9}	2279282	1.70	6.0×10^{-10}	2×10^{-13}	0.34
<i>PPAT</i>	4q12	2×10^{-7}	2030364	1.56	3.1×10^{-6}	8×10^{-5}	0.21
<i>VAMP8</i>	2p11	9×10^{-8}	6547625	1.38	4.8×10^{-5}	2×10^{-10}	0.17
<i>CTSH</i>	15q25	7×10^{-9}	7176604	1.44	1.5×10^{-4}	3×10^{-6}	0.15
<i>IL16</i>	15q26	3×10^{-10}	4128767	1.43	5.0×10^{-4}	0.0029	0.13
<i>ZP3</i>	7q11	9×10^{-10}	306191*	2.70	2.3×10^{-3}	0.0011	0.10
<i>GSTM2</i>	1p13	3×10^{-8}	668413	NA	0.092	>0.5	0.03
<i>TM7SF3</i>	12p11	< 10^{-11}	3134726*	NA	0.42	>0.5	<0.01
<i>ICAP-1A</i>	2p25	< 10^{-11}	434836*	NA	0.79	>0.5	<0.01

*ABI hCV SNP identifier.

†Ratio of mean expression levels of homozygotes for SNPs. NA indicates that the regression was not significant.

‡ P -value for regression of expression level on genotype.

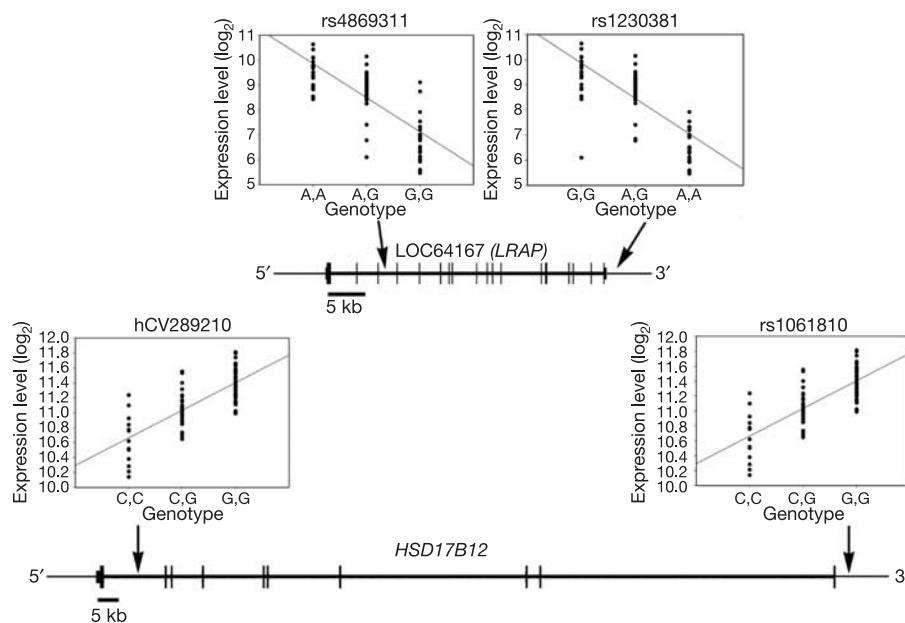


Figure 3 Regression of expression phenotype of LOC64167 and *HSD17B12* on nearby SNPs. For LOC64167, the distance between SNP markers rs4869311 and rs1230381 is 30 kb. For *HSD17B12*, the distance between hCV289210 and rs1061810 is 172 kb.

linkage strategy allowed us to identify numerous transcriptional regulatory loci, without any prior knowledge of the regulatory mechanism. It uncovered a complex network of regulation that includes determinants that influence expression of nearby genes (*cis*-acting), determinants located on other chromosomes (*trans*-acting), and hotspots of genetic determinants that affect the expression of many genes. The approach is reliable and accurate; results from association and differential allelic expression support the linkages *in cis*, suggesting that findings of linkages *in trans* are also valid. Our approach detects differential allelic expression without requiring that sequence variants be located in coding regions.

Many studies have shown that gene expression levels differ according to developmental stages, health and disease, and physiological or other biologically relevant states. However, little is known about natural variation in human gene expression, especially as a result of germ-line differences. Normal variation in gene expression is likely to account for a substantial part of human variation. It will therefore contribute to differences that are important for understanding essential aspects of human biology, including networks of interacting genetic effects, evolution, and susceptibility to complex diseases.

Mapping quantitative traits and unravelling transcriptional control are challenging, even when applied to one phenotype at a time. In studies of typical quantitative human traits like blood pressure or serum levels of metabolites, strong effects are rarely found. Here we have coupled genomic technologies for expression profiling with genome-wide genetic mapping using SNP markers, and shown that specific chromosomal regions contain germ-line determinants that regulate gene expression. These approaches and results show how it will be possible to dissect the genetic contribution to natural variation in human gene expression. □

Methods

CEPH samples and expression phenotyping

The data were from members of 14 CEPH families (CEPH 1333, 1340, 1341, 1345, 1346, 1347, 1362, 1408, 1416, 1418, 1421, 1423, 1424 and 1454). Expression and marker genotype data were available for all parents and a mean of eight offspring per sibship (range 7–9). (Data from grandparents are not used in SIBPAL.)

For the expression analysis, RNA was extracted from lymphoblastoid cells of each individual and hybridized onto Affymetrix Genome Focus Arrays per the manufacturer's protocol. Expression intensity was scaled to 500 and transformed by $\log_2$.

Genotypes

SNP genotypes for 2,756 autosomal SNP markers for individuals whose lymphoblastoid cells were phenotyped were downloaded from The SNP Consortium database of the SNP Consortium Linkage Map Project (http://snp.cshl.org/linkage_maps/). Most SNP Consortium SNPs are clustered in very closely linked sets (two or three SNPs within 100 kb) with average intercluster distance approximately 3 Mb. We used PedStat²¹ to check for mendelian inconsistencies. This resulted in the removal of 815 genotypes at 237 distinct SNP markers.

Analysis of linkage and association

Multipoint genome-wide linkage analysis was done by SIBPAL in S.A.G.E.¹¹. We used the recommended option ('W4' SIBPAL) for weighting pairwise phenotypic differences between siblings²². SIBPAL determines evidence for linkage at each SNP from regression of the phenotype difference between siblings on the estimated proportion of marker alleles shared identical-by-descent between siblings; the result is reported as a *t*-value with corresponding significance, as given in the text. Point-wise significance was converted to genome-wide significance by use of the expression in ref. 13 (see <http://www.imbs.muh-luebeck.de/pub/silcLOD/>). In permutation analysis by S.A.G.E. of the results for the 142 phenotypes ($t > 5$, $P < 4.3 \times 10^{-7}$), we found one phenotype with one *t*-value > 5 among 1,000 replicates; 100,000 additional permutations of this phenotype yielded two more *t*-values > 5 . Further testing with 100,000 replicates for eight phenotypes (three with $4.1 \times 10^{-7} < P < 4.3 \times 10^{-7}$) yielded no *t*-values > 5 .

For association analysis, the $\log_2$ -transformed expression level of 94 unrelated individuals (CEPH grandparents), as the dependent variable, was regressed on SNP genotype (coded 0, 1, 2). Conventional analysis of linear regression was carried out. R^2 was estimated for each phenotype/SNP combination as the ratio of regression sum of squares to total sum of squares.

Master regulator probability

The autosomal genome was divided into 491 windows of 5 Mb each (with smaller windows at the ends of chromosomes). For each of 142 phenotypes, we considered all

SNPs with $t > 5$, $P < 4.3 \times 10^{-7}$. Any window with one or more such SNP was counted as having one 'hit' for that phenotype. Some phenotypes have more than one hit in the genome because some have multiple linkage peaks, or because peaks for some phenotypes are broad and span adjacent 5-Mb windows, in which case, each window is counted as having one hit. There were 318 hits defined this way, representing linkages for the 142 phenotypes. We assumed that if the hits were distributed randomly across the genome, their distribution over windows would be approximately Poisson, with mean 0.65 (318 out of 491).

Clustering

The similarity of the expression phenotypes that mapped to the hotspots of transcriptional control on chromosomes 14 and 20 was assessed by Pearson's correlation (absolute value), and the phenotypes were grouped by hierarchical clustering using the average linkage method. The significance of the correlation between genes was assessed by permutation. For each permutation, the expression levels of the 3,554 genes for each individual were permuted, and all $3,554 \times 3,553/2$ pairwise correlations were calculated. Among the 1,000 permuted sets, the highest pairwise correlation coefficient was 0.52.

SNP genotyping and allele-specific RT-PCR

SNPs in the region of the genes with *cis* regulators were identified using NCBI dbSNP or Applied Biosystems (ABI)/Celera Discovery System. DNA samples were genotyped using ABI TaqMan technology and the 7900 HT Sequence Detection System. PCR was carried out with primer and probe sets (ABI Assay-on-Demand and Assay-by-Design) according to the manufacturer's protocols. Allele-specific RT-PCR was performed using similar protocols with complementary DNA samples as template.

Received 22 March; accepted 5 July 2004; doi:10.1038/nature02797.

Published online 21 July 2004.

- Oleksiak, M. F., Churchill, G. A. & Crawford, D. L. Variation in gene expression within and among natural populations. *Nature Genet.* **32**, 261–266 (2002).
- Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
- Yvert, G. *et al.* Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **3**, 57–64 (2003).
- Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B. & Kinzler, K. W. Allelic variation in human gene expression. *Science* **297**, 1143 (2002).
- Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
- Cheung, V. G. *et al.* Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genet.* **33**, 422–425 (2003).
- Cheung, V. G. & Spielman, R. S. The genetics of variation in gene expression. *Nature Genet.* **32**, 522–525 (2002).
- Cheung, V. G. *et al.* Genetics of quantitative variation in human gene expression. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 403–407 (2003).
- Dausset, J. *et al.* Centre d'étude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577 (1990).
- Matisse, T. C. *et al.* A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *Am. J. Hum. Genet.* **73**, 271–284 (2003).
- S.A.G.E. Statistical Analysis for Genetic Epidemiology. (Statistical Solutions Ltd, Cork, Ireland, 2003).
- Haseman, J. K. & Elston, R. C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19 (1972).
- Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.* **11**, 241–247 (1995).
- Nobrega, M., Ovcharenko, I., Afzal, V. & Rubin, E. Scanning human gene deserts for long-range enhancers. *Science* **302**, 413 (2003).
- Lettice, L. A. *et al.* Disruption of a long-range *cis*-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99**, 7548–7553 (2002).
- Cohen, B. A., Mitra, R. D., Hughes, J. D. & Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.* **26**, 183–186 (2000).
- Caron, H. *et al.* The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science* **291**, 1289–1292 (2001).
- Spellman, P. T. & Rubin, G. M. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J. Biol.* **1**, 5 (2002).
- Abecasis, G. R., Cardon, L. R. & Cookson, W. O. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
- McKenzie, C. A. *et al.* Trans-ethnic fine mapping of a quantitative trait locus for circulating angiotensin I-converting enzyme (ACE). *Hum. Mol. Genet.* **10**, 1077–1084 (2001).
- Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet.* **30**, 97–101 (2002).
- Shete, S., Jacobs, K. B. & Elston, R. C. Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: weighting sums and differences. *Hum. Hered.* **55**, 79–85 (2003).

Acknowledgements We thank T. Matisse and W. Ewens for discussions and advice, and J. Burdick for technical help. Some analyses for this paper were carried out by using the program package S.A.G.E., which is supported by a grant from the National Center for Research Resources. This work is supported by grants from the National Institutes of Health (to R.S.S. and V.G.C.) and the W.W. Smith Endowed Chair (to V.G.C.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.G.C. (vcheung@mail.med.upenn.edu) or R.S.S. (spielman@pobox.upenn.edu). The GEO accession number for the microarray data is GSE1485.

The principle of temperature-dependent gating in cold- and heat-sensitive TRP channels

Thomas Voets¹, Guy Droogmans¹, Ulrich Wissenbach², Annelies Janssens¹, Veit Flockerzi² & Bernd Nilius¹

¹Laboratory of Physiology, Campus Gasthuisberg, KU Leuven, B-3000 Leuven, Belgium

²Experimentelle und Klinische Pharmakologie und Toxikologie, Medizinische Fakultät, Universität des Saarlandes, D-66421 Homburg, Germany

The mammalian sensory system is capable of discriminating thermal stimuli ranging from noxious cold to noxious heat. Principal temperature sensors belong to the TRP cation channel family, but the mechanisms underlying the marked temperature sensitivity of opening and closing ('gating') of these channels are unknown. Here we show that temperature sensing is tightly linked to voltage-dependent gating in the cold-sensitive channel TRPM8 and the heat-sensitive channel TRPV1. Both channels are activated upon depolarization, and changes in temperature result in graded shifts of their voltage-dependent activation curves. The chemical agonists menthol (TRPM8) and capsaicin (TRPV1) function as gating modifiers, shifting activation curves towards physiological membrane potentials. Kinetic analysis of gating at different temperatures indicates that temperature sensitivity in TRPM8 and TRPV1 arises from a tenfold difference in the activation energies associated with voltage-dependent opening and closing. Our results suggest a simple unifying principle that explains both cold and heat sensitivity in TRP channels.

Mammals sense ambient temperature through primary afferent sensory neurons of the dorsal root and trigeminal ganglia^{1,2}. These cells convey thermal information from peripheral tissues to the spinal cord and brain, where the signals are integrated and interpreted, resulting in appropriate reflexive and cognitive responses. The mammalian sensory system is capable of detecting and discriminating thermal stimuli over a broad temperature spectrum, ranging from noxious cold (<8 °C) to noxious heat (>52 °C), which implies the existence of different types of temperature sensors with distinct thermal sensitivities². Accumulated evidence suggests that the principal temperature sensors in the sensory nerve endings of mammals belong to the transient receptor potential (TRP) superfamily of cation channels^{3,4}. At present, six temperature-sensitive TRP channels (or thermoTRPs)² have been described, that together cover almost the entire range of temperatures that mammals are able to sense. Four TRP channels belonging to the TRPV subfamily are activated by heating, with characteristic activation temperatures ranging from warm temperatures (>25 °C for TRPV4; >31 °C for TRPV3)^{5–9} to heat (>43 °C for TRPV1)¹⁰ and noxious heat (>52 °C for TRPV2)¹¹. TRPM8 and TRPA1 (ANKTM1) are activated by cooling, (<28 °C for TRPM8; <18 °C for TRPA1)^{12–14}, although the cold-sensitivity of TRPA1 has been disputed¹⁵.

The origin of the remarkably steep temperature sensitivity of the thermoTRPs is still obscure. Until now, three possible mechanisms for temperature-dependent channel gating have been envisaged³. Changes in temperature could lead to the production and binding of channel-activating ligands. Alternatively, the channel protein may undergo temperature-dependent structural rearrangements leading to channel opening. Finally, thermoTRPs may be able to sense changes in membrane tension due to temperature-dependent lipid bilayer rearrangements. However, direct experimental evidence supporting any of these mechanisms is currently lacking.

TRP channels were originally considered as cation channels with little or no voltage sensitivity. This view was in line with the paucity of positively charged residues in the fourth transmembrane domain (S4) (ref. 3), which is known to function as (part of) the voltage sensor in classical voltage-gated K⁺, Na⁺ and Ca²⁺ channels¹⁶. Nevertheless, recent studies have demonstrated that TRPM4 and

TRPM5, two closely homologous monovalent cation channels, are dually gated by intracellular Ca²⁺ and membrane depolarization^{17–19}. Given that the cold- and menthol-sensitive TRPM8 is closely related to TRPM4 and TRPM5, with more than 40% similarity at the amino acid level, we investigated whether it displays similar voltage dependence and whether this might be related to its cold sensitivity.

Cold activation of TRPM8

In whole-cell patch-clamp experiments, cooling to 15 °C (Fig. 1a) or application of the cooling agent menthol (Fig. 1b) activated a robust current in human embryonic kidney (HEK) 293 cells transiently transfected with TRPM8, in line with previous reports^{12,13}. Non-transfected or vector-transfected cells were unresponsive to either stimulus. Activation of macroscopic TRPM8 currents by cold and menthol also occurred in cell-free inside-out patches (Fig. 1c, d; data not shown), demonstrating that these stimuli function in a membrane-delimited manner. It should be noted that the temperature-sensitivity of TRPM8 in inside-out patches was shifted towards lower temperatures when compared to whole-cell measurements, which could be due to the loss of a regulatory factor.

A typical feature of TRPM8 currents is the pronounced outward rectification (Fig. 1a–d). Rectification could either be an intrinsic property of the pore or alternatively reflect a voltage-dependent mechanism that closes the channel at negative potentials. Evidence for the latter mechanism was obtained using a classical tail current protocol (Fig. 1e): TRPM8 activates upon depolarization to +120 mV and rapidly closes at negative voltages. Notably, the current–voltage relation obtained immediately after the prepulse to +120 mV was linear (Fig. 1f), indicating that open TRPM8 channels display an ohmic current–voltage relation. The time-dependent closure of TRPM8 at negative potentials persisted when Ca²⁺ and Mg²⁺ were omitted from the intra- and extracellular solutions, excluding block by divalent cations as a cause of rectification (data not shown). Taken together, our data demonstrate that TRPM8 is a voltage-dependent channel activated by membrane depolarization. Outward rectification arises from the rapid and voltage-dependent closure of the channel at negative voltages.

Next, we asked how the voltage dependence of TRPM8 relates to its function as a cold receptor^{12,13}. To address this, we tested whether

membrane voltage influences the cold sensitivity of the channel by measuring current activation at different holding potentials during slow cooling of TRPM8-expressing cells. During such cooling protocols we consistently found that current activation at depolarized potentials precedes that at more negative potentials (Fig. 2a, b). At +100 mV, robust outward currents were already activated at 32 °C ($1,080 \pm 230$ pA ($n = 12$) versus 69 ± 20 pA ($n = 10$) in non-transfected cells; $P < 0.001$), whereas inward currents at -80 mV were only discernible upon cooling below ~ 25 °C (Fig. 2a, b). Thus, the cold sensitivity of TRPM8 strongly depends on the transmembrane voltage.

Using a voltage step protocol ranging from -120 to $+160$ mV (and in some cases up to $+220$ mV), we determined how temperature affects the voltage dependence of TRPM8 (Fig. 2c). At 37 °C, significant outward currents could only be measured at potentials above $+100$ mV, and the midpoint voltage of activation ($V_{1/2}$) was around $+200$ mV (Fig. 2c–e). Cooling induced a leftward shift of the activation curve resulting in channel activity at more physio-

logical voltages: $V_{1/2}$ decreased by approximately 150 mV upon cooling from 37 to 15 °C (7.3 mV °C $^{-1}$), and saturated around $+25$ mV between 5 and 10 °C (Fig. 2e). From this we conclude that cooling activates TRPM8 by causing a drastic shift of the voltage dependence of activation.

Heat activation of TRPV1

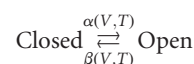
To test whether a comparable process underlies temperature-dependent gating in a TRP channel with opposite temperature sensitivity, we performed similar experiments in HEK 293 cells expressing TRPV1 (ref. 10). Although TRPV1 was originally described as voltage-independent¹⁰, some voltage-dependent properties of the channel have been recently described^{20,21}. During slow heating ramps, we observed that activation of TRPV1 at depolarized voltages occurred at significantly lower temperatures than at hyperpolarized voltages (Fig. 3a, b): at -100 mV, TRPV1 currents only became prominent at temperatures above 40 °C, whereas at $+100$ mV robust outward currents were already present at room temperature ($1,742 \pm 279$ pA ($n = 17$) versus 59 ± 10 pA ($n = 10$) in non-transfected cells; $P < 0.001$). Thus, similar to TRPM8, the temperature sensitivity of TRPV1 is strongly dependent on the transmembrane voltage.

We then determined how temperature affects the voltage dependence of TRPV1. At 17 °C, significant outward currents could only be measured at potentials above 100 mV, and $V_{1/2}$ was around $+150$ mV (Fig. 3c–e). Heating induced a gradual leftward shift of the activation curve, resulting in channel activity at more physiological voltages: $V_{1/2}$ decreased by approximately 200 mV upon heating from 17 to 40 °C (9.1 mV °C $^{-1}$), reaching a value of -50 mV between 40 and 45 °C (Fig. 3e). Thus, heat-induced activation of TRPV1 is the result of a drastic leftward shift of the voltage dependence of activation.

The principle of temperature sensing

Our present results have important implications when considering possible mechanisms for temperature sensing in TRP channels³. First, we found that TRPM8 is activated by cooling in cell-free patches. Similarly, previous work has demonstrated heat-activation of TRPV1 in cell-free patches²². These results imply that temperature sensing in both channels is a membrane-delimited process, and argue against temperature-dependent binding of second messengers as a mechanism for channel activation. Second, our results demonstrate that temperature-dependent activation of TRPM8 and TRPV1 is not governed by a single characteristic thermal threshold. Temperature sensitivity is modulated by the transmembrane voltage, and changes in ambient temperature result in graded shifts of the voltage dependence of channel activation. These results are not in line with the suggestion that channel activation results from a temperature-dependent phase transition of the lipid membrane or a conformational transition (or denaturation) of the channel protein structure³, as such processes would predict a single sharp thermal threshold²³.

As an alternative mechanism, we explored whether temperature sensitivity could be the thermodynamic consequence of differences in the activation energies associated with voltage-dependent channel opening and closing. Given that the time course of TRPM8 and TRPV1 currents during voltage steps was mostly well described with a single exponential function, we employed the simplest, two-state model for a voltage-gated channel:



The opening and closing rates α and β are related to membrane voltage and temperature according to^{16,24}

$$\alpha = A e^{-\frac{E_{a,open}}{RT}} e^{\frac{\delta z F V}{RT}} \quad \text{and} \quad \beta = B e^{-\frac{E_{a,close}}{RT}} e^{\frac{-(1-\delta) z F V}{RT}} \quad (1)$$

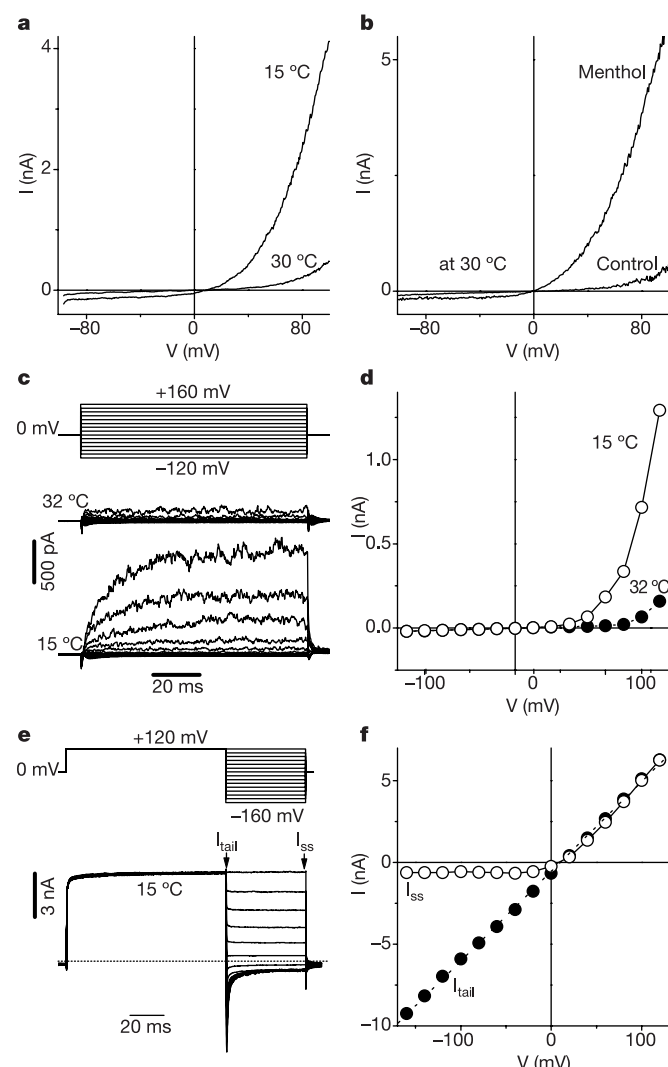


Figure 1 Outward rectification of TRPM8 is due to voltage-dependent gating. **a, b**, Current–voltage (I – V) relations of whole-cell TRPM8 currents activated by cooling (**a**) or application of $100 \mu\text{M}$ menthol (**b**). Currents were measured during 200-ms voltage ramps from -100 to $+100$ mV. **c**, Effect of cooling on TRPM8-currents measured in inside-out patches. **d**, I – V relations obtained at the end of the voltage-steps shown in **c**. **e**, Whole-cell TRPM8 currents at 15 °C in response to the indicated voltage protocol. **f**, Tail (I_{tail}) and steady-state (I_{ss}) currents obtained from the voltage-steps shown in **e**. The dashed line represents a linear fit.

where $E_{a,open}$ and $E_{a,close}$ represent the activation energies associated with channel opening and closing, R the gas constant ($8.31 \text{ J mol}^{-1} \text{ K}^{-1}$), T the absolute temperature, z the effective charge associated with voltage-dependent gating, δ is the fraction of z moved in the outward direction, F the Faraday constant ($9.65 \times 10^4 \text{ C mol}^{-1}$), V the transmembrane voltage and A and B are preexponential factors. According to Eyring rate theory, activation energy and preexponential factor are related to the enthalpic (ΔH) and entropic (ΔS) component of the activation barrier, according to $E_a = \Delta H + RT$ and $A = k_0 e^{\Delta S/R}$ (where k_0 is the frequency factor)^{16,24}. Experimental values for z were obtained by fitting the Boltzmann function to the activation curves (Figs 2d and 3d). Opening and closing rates were obtained from the mono-exponential time constant τ of current activation/deactivation and the steady-state P_{open} using the expressions $\alpha = P_{open}/\tau$ and $\beta = (1 - P_{open})/\tau$ (see Supplementary Fig. S1). Values for α and β as a function of temperature were then displayed in Arrhenius

plots (that is, $\log(\text{rate})$ versus $1/T$), which allows determination of the activation energy from the slope of the curves.

For TRPM8, the opening rate α was characterized by an $E_{a,open}$ value of 15.7 kJ mol^{-1} corresponding to a 10°C temperature coefficient (Q_{10}) of 1.2, indicative of a shallow temperature dependence (Fig. 4a). In contrast, the closing rate β was steeply temperature-dependent, reflected in a ~ 10 times higher value for $E_{a,close}$ of 173 kJ mol^{-1} , which corresponds to a Q_{10} of 9.4 (Fig. 4b). Estimated activation energies were not significantly altered by membrane voltage, which can be appreciated from the parallel Arrhenius plots at different potentials (Fig. 4a, b). Subsequently, we simulated TRPM8 currents using the two-state gating model with the experimentally deduced parameters. Figure 4c shows the simulated temperature dependence of TRPM8 currents at two potentials in response to the cooling ramp used in the experiment of Fig. 2a. The model adequately predicts the cold-induced current activation and the difference in temperature sensitivity at -80 and $+100 \text{ mV}$

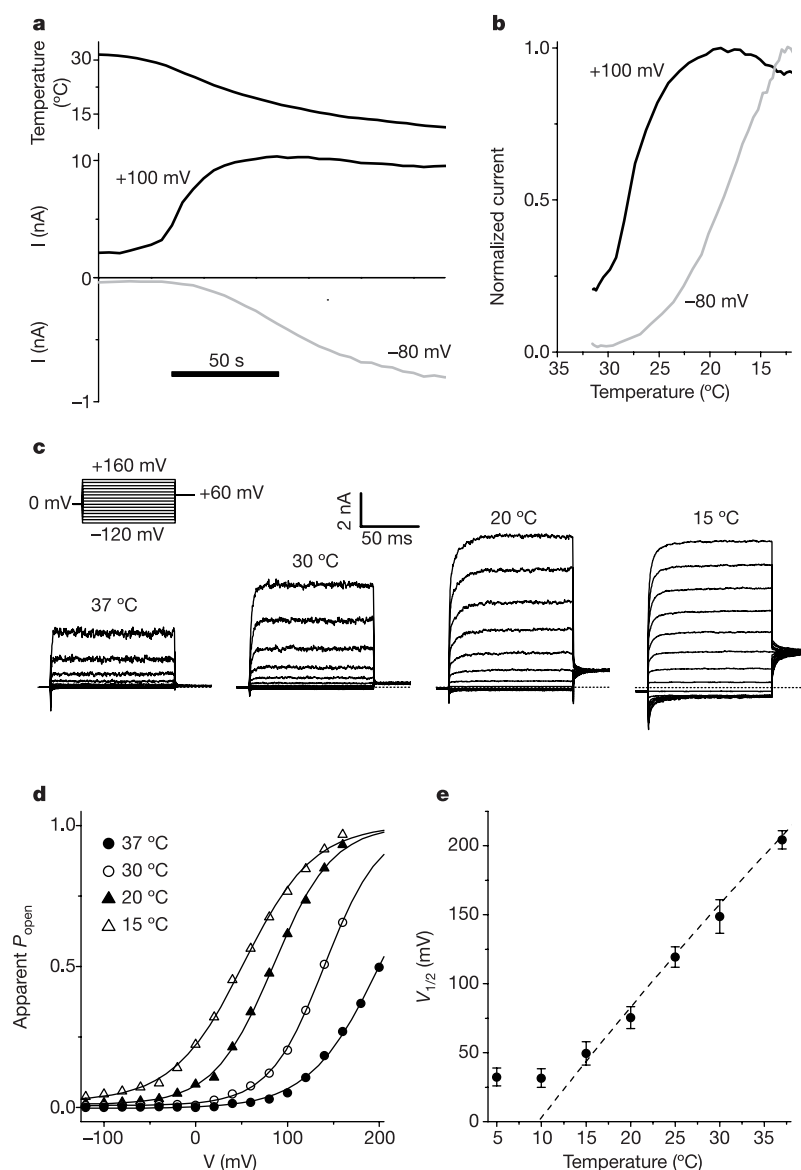


Figure 2 Cooling activates TRPM8 by shifting the voltage dependence of activation. **a**, Whole-cell TRPM8 currents at +100 and -80 mV in response to slow cooling of the bath solution. **b**, Normalized current responses at +100 and -80 mV in function of temperature. **c**, Current traces at different temperatures in response to the indicated voltage protocol. **d**, Steady-state activation curves at different temperatures for the

currents shown in **c**. The apparent P_{open} in function of voltage was determined as explained in the methods section. Lines represent Boltzmann functions fitted to the data. **e**, $V_{1/2}$ as a function of temperature (n values between 6 and 15). Dashed line represents the change in $V_{1/2}$ as predicted by the two-state model.

(compare Figs 2b and 4c). Likewise, simulated TRPM8 currents in response to voltage steps at different temperatures (Fig. 4d) were strikingly similar to the experimental data (Fig. 2c). Finally, the model predicted the drastic leftward shift of $V_{1/2}$ upon cooling (dashed line in Fig. 2e), except for a small deviation at temperatures below 10 °C, which might reflect channel desensitisation.

In the case of TRPV1, the temperature dependencies of α and β were opposite to those of TRPM8 (Fig. 4e, f): the opening rate α was steeply temperature-dependent ($E_{a,open} = 208 \text{ kJ mol}^{-1}$; $Q_{10} = 14.8$), whereas the closing rate β showed a shallow temperature dependence ($E_{a,close} = 23.2 \text{ kJ mol}^{-1}$; $Q_{10} = 1.35$). Using the two-state model with the experimentally deduced parameters, we could faithfully reproduce the temperature dependence of TRPV1 current at different potentials (Fig. 4g), TRPV1 currents in response to voltage steps at different temperatures (Fig. 4h) and the shift of $V_{1/2}$ as a function of temperature (dashed line in Fig. 3e).

The excellent match between model simulations and experimental results for both TRPM8 and TRPV1 led us to formulate a general

principle for temperature sensitivity. Temperature sensitivity occurs whenever the activation energies associated with the opening and closing transitions are sufficiently different. When $E_{a,open} \ll E_{a,close}$, the open probability of the channel will increase upon cooling, as illustrated by the behaviour of TRPM8 (Fig. 4a–d). On the contrary, the open probability of the channel will increase upon heating when $E_{a,open} \gg E_{a,close}$, as is the case for TRPV1 (Fig. 4e–h). In most voltage-gated K^+ , Ca^{2+} and Na^+ channels the temperature dependence of channel activation and deactivation are similar ($E_{a,open} \approx E_{a,close}$)^{16,25,26}. In these channels, increasing temperature leads to an acceleration of channel gating, without major changes in the steady-state open probability.

Although this principle is rather fundamental, we do not exclude that other thermoTRPs exploit different mechanisms. For example, heat activation of TRPV4 does not occur in cell-free patches^{5,27}, possibly reflecting the necessity of a diffusible messenger. We would also like to point out that a simple two-state model is probably an oversimplification of the full gating intricacies of TRPV1 and

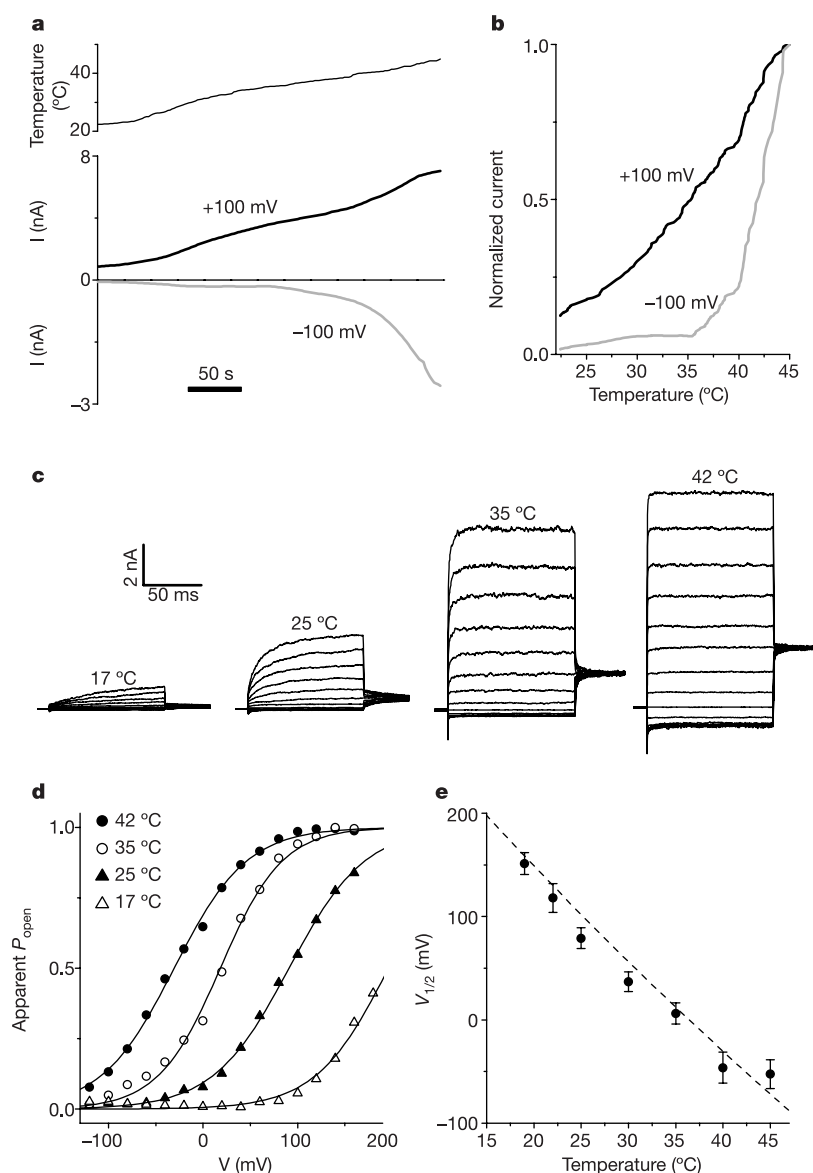


Figure 3 Heating activates TRPV1 by shifting the voltage dependence of activation. **a**, Whole-cell TRPV1 currents at +100 and -100 mV in response to slow heating of the bath solution. **b**, Normalized current responses at +100 and -100 mV as a function of temperature. **c**, Current traces at different temperatures in response to voltage steps

ranging from -120 to +160 mV. **d**, Steady-state activation curves at different temperatures for the currents shown in **c**. Lines represent Boltzmann functions fitted to the data. **e**, $V_{1/2}$ as a function of temperature (n values between 4 and 9). Dashed line represents the change in $V_{1/2}$ as predicted by the two-state model (see text).

TRPM8. Indeed, single-channel measurements of TRPV1 provided evidence for multiple open and closed states²⁸. Likewise, we observed that a single exponential function was not always optimal for description of the time dependence of TRPM8, suggesting the presence of more than one open and/or closed state. Nevertheless, the two-state model represents a good approximation of multi-state models as long as the transition with the steepest temperature dependence is rate limiting.

Effect of ligand activators

TRPM8 and TRPV1 not only function as temperature sensors, but also as ionotropic receptors for a variety of chemical substances. TRPV1 is directly gated by several chemicals that cause a burning sensation such as vanilloid compounds and protons^{10,22,29}, as well as by the endocannabinoid anandamide³⁰. In contrast, TRPM8 is activated by several plant-derived and synthetic cooling compounds^{12,13,31}. Our present results raised the question whether

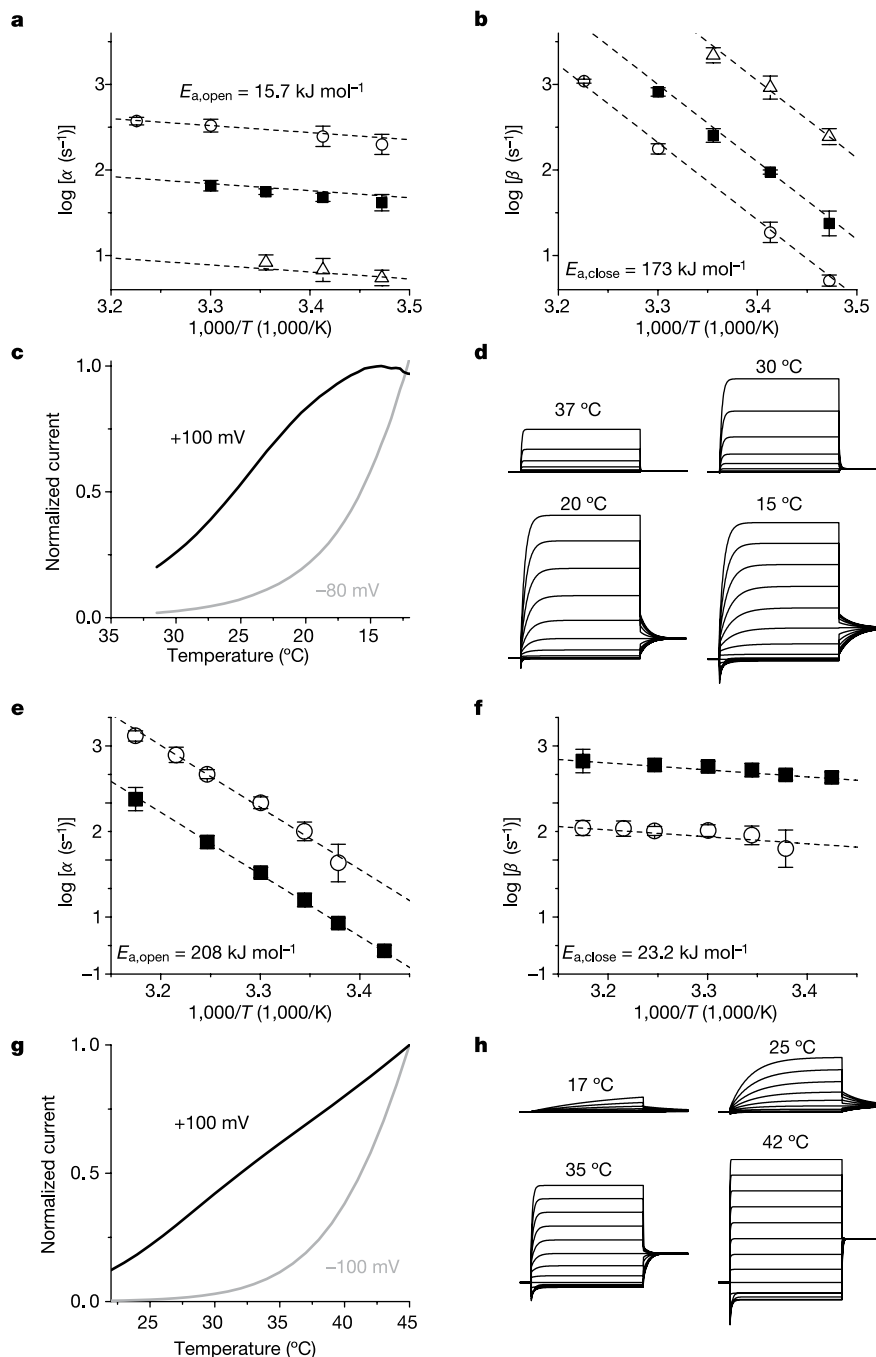


Figure 4 A two-state model accurately predicts the temperature-dependent behaviour of TRPM8 and TRPV1. **a, b**, Arrhenius plots of α and β for TRPM8 measured at +160 mV (open circles), +60 mV (filled squares) or -80 mV (open triangles). We obtained the following parameters: $A = 1.37 \times 10^4 \text{ s}^{-1}$, $E_{a,\text{open}} = 15.7 \text{ kJ mol}^{-1}$, $B = 1.94 \times 10^{33} \text{ s}^{-1}$, $E_{a,\text{close}} = 173 \text{ kJ mol}^{-1}$, $\delta = 0.5$ and $z = 0.82$. **c**, Simulated TRPM8 current at +100 and -80 mV as a function of temperature. **d**, Simulated TRPM8

currents in response to voltage steps (protocol as in Fig. 2c). **e, f**, Arrhenius plots of α and β for TRPV1 measured at +100 (open circles) and -100 mV (filled squares). We obtained the following parameters: $A = 1.61 \times 10^{37} \text{ s}^{-1}$, $E_{a,\text{open}} = 208 \text{ kJ mol}^{-1}$, $B = 9.67 \times 10^{55} \text{ s}^{-1}$, $E_{a,\text{close}} = 23.2 \text{ kJ mol}^{-1}$, $\delta = 0.5$ and $z = 0.71$. **g**, Simulated TRPV1 currents at +100 and -100 mV as a function of temperature. **h**, Simulated TRPV1 currents in response to voltage steps.

these agonists function by modifying the voltage-dependent properties of these channels. To investigate this, we focused on the two best-studied agonists, namely menthol (TRPM8) and capsaicin (TRPV1).

Analogous to previous work^{12,13}, we found that a low dose of menthol (30 μ M) has little effect on inward TRPM8 current at

34 $^{\circ}$ C, but significantly shifts the cold sensitivity of the channel to higher temperatures (Fig. 5a). The voltage step protocol showed that 30 μ M menthol caused a strong potentiation of outward currents, a prominent slowing of tail current deactivation and a significant leftward shift of the activation curve (Fig. 5b, c). The menthol-induced leftward shift of the activation curve was

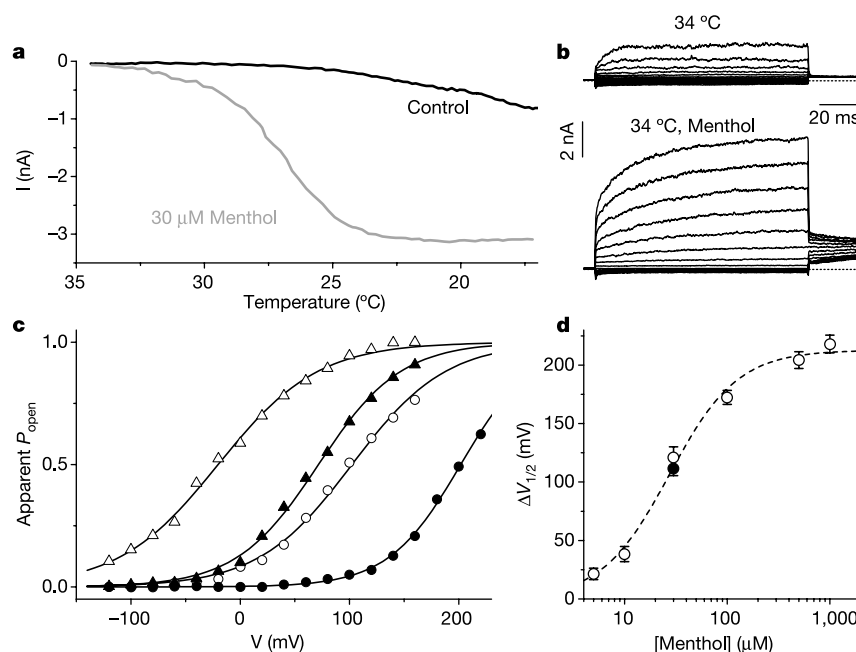


Figure 5 Menthol shifts the TRPM8 activation curve. **a**, TRPM8 currents at -60 mV in response to slow cooling of the bath solution in the absence and presence of 30 μ M menthol. **b**, Currents at 34 $^{\circ}$ C in response to voltage steps ranging from -120 to $+160$ mV in the absence and presence of 30 μ M menthol. **c**, Activation curves at 34 $^{\circ}$ C

(filled symbols) and 24 $^{\circ}$ C (open symbols) in the absence (circles) and presence (triangles) of 30 μ M menthol. **d**, Leftward shift of $V_{1/2}$ ($\Delta V_{1/2}$) as a function of menthol concentration at 32–34 $^{\circ}$ C (open circles; n values between 5 and 20) or 22–24 $^{\circ}$ C (filled circle; $n = 6$).

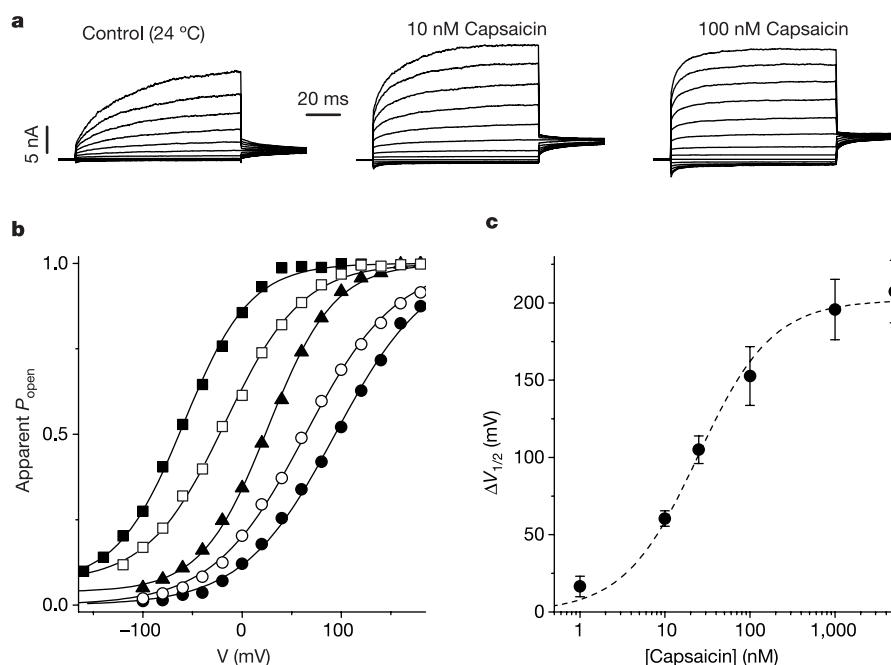


Figure 6 Capsaicin shifts the TRPV1 activation curve. **a**, Currents at 24 $^{\circ}$ C in response to voltage steps ranging from -100 to $+160$ mV at the following capsaicin concentrations: 0 nM (closed circles), 1 nM (open circles), 10 nM (closed triangle), 100 nM (open squares)

and 1 μ M (closed squares). **b**, Steady-state activation curves at 24 $^{\circ}$ C at the indicated capsaicin concentrations. **c**, $\Delta V_{1/2}$ as a function of capsaicin concentration at 24–26 $^{\circ}$ C ($n = 4$ –10).

independent of temperature (Fig. 5c, d), which explains the enhanced cold sensitivity of TRPM8 in the presence of low doses of the cooling agent. The effect of menthol on the activation curve was dose-dependent, with a half-maximal change in $V_{1/2}$ at 27.4 μ M (Fig. 5d). Maximal shifts induced by menthol (218 mV at 1 mM; Fig. 5d) were larger than those induced by cooling to 5 °C (Fig. 2e), in line with previous findings showing that menthol is a more efficacious TRPM8 agonist than cold¹².

Capsaicin, the main pungent ingredient of 'hot' chilli peppers, is one of the most potent chemical activators of TRPV1^{10,22}. We found that submicromolar concentrations of capsaicin had a robust effect on outward TRPV1 currents at 24 °C, and induced a clear leftward shift of the activation curve (Fig. 6a, b). The effect of capsaicin on the TRPV1 activation curve was dose-dependent, with a half-maximal shift at 28.5 nM and a maximal change in $V_{1/2}$ of ~200 mV (Fig. 6b, c). Note that significant channel desensitisation occurs at capsaicin concentrations >100 nM, which may lead to an underestimation of the shift in $V_{1/2}$ at the highest concentrations tested.

Conclusions

Our results show an unexpected tight link between temperature sensing and voltage-dependent gating in two thermoTRPs with opposite temperature sensitivity. Cold activation of TRPM8 and heat activation of TRPV1 are well described by a single thermodynamic principle, whereby thermosensitivity arises from the difference in activation energies associated with voltage-dependent opening and closing. Chemical agonists of these thermoTRPs function as gating modifiers to mimic and potentiate the thermal responses. At present, the structural determinants of voltage sensing in TRP channels are not known, but the interplay between voltage and temperature sensing implies that any mutational manipulation to the voltage sensor will lead to altered temperature sensitivity. Finally, in a physiological context, our results suggest that the membrane voltage contributes to the fine-tuning of cold- and heat-sensitivity in sensory cells. □

Methods

Electrophysiology

HEK293 were grown in DMEM containing 10% (v/v) fetal calf serum, 4 mM L-alanyl-L-glutamine, 100 U ml⁻¹ penicillin and 100 μ g ml⁻¹ streptomycin at 37 °C in a humidity controlled incubator with 10% CO₂. Cells were transiently transfected with human TRPM8 or human TRPV1 cloned in the bicistronic pCAGGS-IRES-GFP vector using TransIT-293 transfection reagent (Mirus). Between 16 and 24 h after transfection, currents were recorded in the whole-cell or inside-out configuration of the patch-clamp technique using an EPC-9 amplifier and PULSE software (HEKA Elektronik).

Data were sampled at 5–20 kHz and digitally filtered off-line at 1–5 kHz. Between 60% and 90% of the series resistance was compensated, reducing voltage errors to less than 10 mV. The standard intracellular solution for whole-cell measurements contained: 150 mM NaCl, 3 mM MgCl₂, 5 mM EGTA and 10 mM HEPES at pH 7.2. The standard extracellular solution contained: 150 mM NaCl, 6 mM CsCl, 1 mM MgCl₂, 1.5 mM CaCl₂, 10 mM glucose and 10 mM HEPES at pH 7.4. In some experiments, a divalent cation-free extracellular solution was used, which contained 150 mM NaCl, 10 mM EDTA, 10 mM glucose and 10 mM HEPES at pH 7.4. For inside-out measurements, we used the standard extracellular solution in the pipette and the standard intracellular solution in the bath. The temperature of the perfusate was controlled using a SC-20 dual in-line heater/cooler (Warner Instruments). Reported temperature was measured using a TA-29 Thermistor (Thermometrics) placed within 500 μ m of the patch-clamped cell.

Data analysis and model simulation

Data analysis, model simulations and data display was performed using Origin 6.1 (OriginLab Corporation) or Igor Pro 4.0 (Wavemetrics). Group data are expressed as mean $\pm$ s.e.m. from n cells. The Student's unpaired t -test was used for statistical comparison. Two distinct methods were used to obtain activation curves using the voltage protocol shown in Figs 2c and 3c. In the first approach, tail currents were measured during the first millisecond of the final step +60 mV and normalized to the maximal tail current. However, tail current activation/deactivation at +60 mV was often too fast (time constants <1 ms) for an accurate determination of P_{open} , especially at higher temperatures. In a second approach, steady-state currents at the end of the voltage steps were divided by the estimated fully-activated current-voltage relation of the channel.

These open channel current-voltage relations were deduced from tail currents following strongly depolarizing prepulses, similar to the one shown in Fig. 1e, f. In those cases where both approaches were feasible, estimates for $V_{1/2}$ differed by less than 10 mV. For model simulations, we used analytical expressions of P_{open} as a function of temperature, voltage and time for the two-state model. Currents were then calculated assuming a linear conductance and a Q_{10} value for the single-channel conductance of 1.35, similar to that of aqueous diffusion¹⁶.

Received 31 March; accepted 7 June 2004; doi:10.1038/nature02732.

- Julius, D. & Basbaum, A. I. Molecular mechanisms of nociception. *Nature* **413**, 203–210 (2001).
- Patapoutian, A., Peier, A. M., Story, G. M. & Viswanath, V. ThermoTRP channels and beyond: mechanisms of temperature sensation. *Nature Rev. Neurosci.* **4**, 529–539 (2003).
- Clapham, D. E. TRP channels as cellular sensors. *Nature* **426**, 517–524 (2003).
- Voets, T. & Nilius, B. TRPs make sense. *J. Membr. Biol.* **192**, 1–8 (2003).
- Watanabe, H. *et al.* Heat-evoked activation of TRPV4 channels in a HEK293 cell expression system and in native mouse aorta endothelial cells. *J. Biol. Chem.* **277**, 47044–47051 (2002).
- Güler, A. D. *et al.* Heat-evoked activation of the ion channel, TRPV4. *J. Neurosci.* **22**, 6408–6414 (2002).
- Xu, H. *et al.* TRPV3 is a calcium-permeable temperature-sensitive cation channel. *Nature* **418**, 181–186 (2002).
- Smith, G. D. *et al.* TRPV3 is a temperature-sensitive vanilloid receptor-like protein. *Nature* **418**, 186–190 (2002).
- Peier, A. M. *et al.* A heat-sensitive TRP channel expressed in keratinocytes. *Science* **296**, 2046–2049 (2002).
- Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
- Caterina, M. J., Rosen, T. A., Tominaga, M., Brake, A. J. & Julius, D. A capsaicin-receptor homologue with a high threshold for noxious heat. *Nature* **398**, 436–441 (1999).
- McKemy, D. D., Neuhauser, W. M. & Julius, D. Identification of a cold receptor reveals a general role for TRP channels in thermosensation. *Nature* **416**, 52–58 (2002).
- Peier, A. M. *et al.* A TRP channel that senses cold stimuli and menthol. *Cell* **108**, 705–715 (2002).
- Story, G. M. *et al.* ANKTM1, a TRP-like channel expressed in nociceptive neurons, is activated by cold temperatures. *Cell* **112**, 819–829 (2003).
- Jordt, S. E. *et al.* Mustard oils and cannabinoids excite sensory nerve fibres through the TRP channel ANKTM1. *Nature* **427**, 260–265 (2004).
- Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, Sunderland, Massachusetts, 2001).
- Launay, P. *et al.* TRPM4 is a Ca²⁺-activated nonselective cation channel mediating cell membrane depolarization. *Cell* **109**, 397–407 (2002).
- Nilius, B. *et al.* Voltage dependence of the Ca²⁺-activated cation channel TRPM4. *J. Biol. Chem.* **278**, 30813–30820 (2003).
- Hofmann, T., Chubonov, V., Gudermann, T. & Montell, C. TRPM5 is a voltage-modulated and Ca²⁺-activated monovalent selective cation channel. *Curr. Biol.* **13**, 1153–1158 (2003).
- Gunthorpe, M. J., Harries, M. H., Prinjha, R. K., Davis, J. B. & Randall, A. Voltage- and time-dependent properties of the recombinant rat vanilloid receptor (rVR1). *J. Physiol. (Lond.)* **525**, 747–759 (2000).
- Vlachova, V. *et al.* Functional role of C-terminal cytoplasmic tail of rat vanilloid receptor 1. *J. Neurosci.* **23**, 1340–1350 (2003).
- Tominaga, M. *et al.* The cloned capsaicin receptor integrates multiple pain-producing stimuli. *Neuron* **21**, 531–543 (1998).
- Marsh, D. General features of phospholipid phase transitions. *Chem. Phys. Lipids* **57**, 109–120 (1991).
- Sigworth, F. J. Voltage gating of ion channels. *Q. Rev. Biophys.* **27**, 1–40 (1994).
- Tiwari, J. K. & Sikdar, S. K. Temperature-dependent conformational changes in a voltage-gated potassium channel. *Eur. Biophys. J.* **28**, 338–345 (1999).
- van Lunteren, E., Elmslie, K. S. & Jones, S. W. Effects of temperature on calcium current of bullfrog sympathetic neurons. *J. Physiol. (Lond.)* **466**, 81–93 (1993).
- Chung, M. K., Lee, H. & Caterina, M. J. Warm temperatures activate TRPV4 in mouse 308 keratinocytes. *J. Biol. Chem.* **278**, 32037–32046 (2003).
- Liu, B., Hui, K. & Qin, F. Thermodynamics of heat activation of single capsaicin ion channels VR1. *Biophys. J.* **85**, 2988–3006 (2003).
- Jordt, S. E., Tominaga, M. & Julius, D. Acid potentiation of the capsaicin receptor determined by a key extracellular site. *Proc. Natl Acad. Sci. USA* **97**, 8134–8139 (2000).
- Zygmunt, P. M. *et al.* Vanilloid receptors on sensory nerves mediate the vasodilator action of anandamide. *Nature* **400**, 452–457 (1999).
- Behrendt, H. J., Germann, T., Gillen, C., Hatt, H. & Jostock, R. Characterization of the mouse cold-menthol receptor TRPM8 and vanilloid receptor type-1 VR1 using a fluorometric imaging plate reader (FLIPR) assay. *Br. J. Pharmacol.* **141**, 737–745 (2004).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Talavera, G. Owsianik, J. Vriens, F. Mahieu and J. Prenen for helpful suggestions and criticisms, and C. Benham (GlaxoSmithKline) for providing us with the human TRPV1 clone. This work was supported by the Belgian Federal Government, the Flemish Government and the Onderzoeksrada KU Leuven. T.V. is a postdoctoral Fellow of the Fund for Scientific Research, Flanders (Belgium) (FWO-Vlaanderen).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T.V. (Thomas.Voets@med.kuleuven.ac.be).

Transport of solar wind into Earth's magnetosphere through rolled-up Kelvin–Helmholtz vortices

H. Hasegawa¹, M. Fujimoto², T.-D. Phan³, H. Rème⁴, A. Balogh⁵, M. W. Dunlop⁶, C. Hashimoto² & R. TanDokoro²

¹Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755-8000, USA

²Department of Earth and Planetary Sciences, Tokyo Institute of Technology, Meguro, Tokyo 152-8551, Japan

³Space Sciences Laboratory, University of California, Berkeley, California 94720-7540, USA

⁴Centre d'Etude Spatiale des Rayonnements, BP 4346, Toulouse 31029, France

⁵Space and Atmospheric Physics Group, Imperial College, London SW7 2BZ, UK

⁶Space Sciences Division, Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 0QX, UK

Establishing the mechanisms by which the solar wind enters Earth's magnetosphere is one of the biggest goals of magnetospheric physics, as it forms the basis of space weather phenomena such as magnetic storms and aurorae¹. It is generally believed that magnetic reconnection is the dominant process, especially during southward solar-wind magnetic field conditions when the solar-wind and geomagnetic fields are antiparallel at the low-latitude magnetopause². But the plasma content in the outer magnetosphere increases during northward solar-wind magnetic field conditions^{3,4}, contrary to expectation if reconnection is dominant. Here we show that during northward solar-wind magnetic field conditions—in the absence of active reconnection at low latitudes—there is a solar-wind transport mechanism associated with the nonlinear phase of the Kelvin–Helmholtz instability⁵. This can supply plasma sources for various space weather phenomena.

Along the outer boundary of Earth's magnetosphere, there is a boundary layer that contains plasma of dominantly solar-wind origin⁶ (Fig. 1a). The boundary layer exists for all orientations of the solar-wind magnetic field, but it tends to be thicker when the solar-wind field points northward^{3,4}. The existence of the boundary layer implies penetration of solar wind across the magnetopause. Although reconnection between solar-wind and terrestrial magnetic fields can readily account for solar-wind entry during southward solar-wind magnetic field conditions, it is at present not known how the plasma crosses the magnetopause when the solar-wind magnetic field is oriented northward and parallel to geomagnetic fields and reconnection is less efficient, although there is a suggestion that simultaneous northern and southern cusp reconnection could result in the formation of the boundary layer at low latitudes⁷. Several candidate local entry mechanisms unrelated to reconnection have been proposed⁸, one of which is the Kelvin–Helmholtz instability (KHI) that could occur along the flanks of the magnetosphere where the shocked solar wind is flowing fast relative to the stagnant magnetospheric plasma^{9,10} (Fig. 1). Recent numerical simulation models^{11–16} suggest that fast plasma transport across the magnetopause can be accomplished by the KHI only when the KHI has grown sufficiently to form rolled-up vortices that can engulf plasmas from both sides of the magnetopause. In these models, the collapse of, or reconnection within, such a vortex (in the nonlinear phase of the KHI) is responsible for the plasma transport.

Multiple and quasi-periodic encounters by spacecraft with the magnetopause and vortex-like flow perturbations near the magnetopause have been reported, and are often interpreted as representing surface waves or vortices excited by the KHI^{17–21}. But as long as

these signatures are observed only by a single spacecraft, one cannot tell unambiguously whether the KHI has reached its nonlinear stage and is generating rolled-up vortices, which are the crucial ingredient for plasma transport, or if they are just ripples or small-amplitude Kelvin–Helmholtz (KH) vortices on the magnetopause surface. The KH rolled-up vortex expected to form at the magnetopause has complex structures, such as vortical plasma flow and a filament-like high density region intruding into the low density (magnetospheric) region (see Fig. 1). To resolve such complex structures in the KH vortices, multipoint *in situ* measurements as carried out by the Cluster mission are essential, as is comparison with realistic three-dimensional (3D) plasma simulations.

Here we report multi-spacecraft measurements that provide unambiguous evidence for rolled-up vortices at the flank magnetopause as well as simultaneously observed boundary-layer characteristics that result from plasma transport across the boundary, such as would be expected from the suggested KHI mechanism.

The four Cluster spacecraft forming a tetrahedron made a

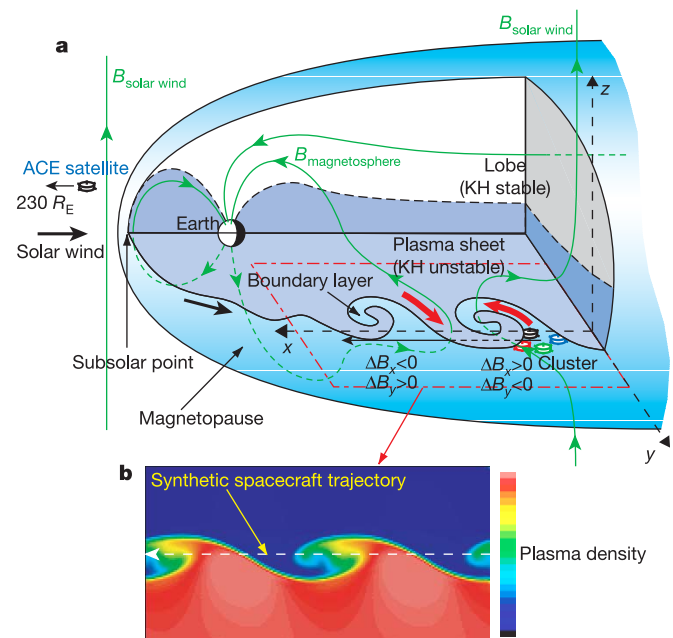


Figure 1 Three-dimensional (3D) cutaway view of Earth's magnetosphere, showing signatures of Kelvin–Helmholtz instability (KHI). **a**, View of the magnetosphere, showing the KH vortices at the duskside magnetopause. The velocity gradient across the magnetopause increases with distance from the subsolar point. The KHI occurs at the interface between the solar wind and the plasma sheet because the plasma energy dominates in both regions, whereas it does not occur at the surface of the lobes where the magnetic energy dominates and the magnetic tension prevents it from deforming the magnetopause. Consequently, the KH vortices evolve only along the low-latitude magnetopause and only low-latitude portions of the magnetospheric and solar-wind field lines are entrained into the vortices, inducing characteristic field perturbations in regions off the equatorial plane where the Cluster spacecraft were located. The satellites were situated at $(x, y, z) \approx (-3, 19, -3)$ Earth radii (R_E) in GSM (geocentric solar magnetospheric) coordinates, and were separated by $\sim 2,000$ km from each other. The coordinate system in the figures is defined such that $-x$ is in the direction of motion of the vortex structure which was sliding anti-sunward in the spacecraft frame, y points outward along the magnetopause normal, and z points to the north. **b**, Vortex structure resulting from a 3D numerical simulation of the magnetohydrodynamic KHI under a magnetosphere-like geometry, with the plasma sheet sandwiched between the two lobes. Colour-coded is the plasma density (minimum, 0 cm^{-3} ; maximum, 5.0 cm^{-3}) in an x - y cross-section cut below the equatorial plane. The density, velocity and magnetic field variations expected when a synthetic satellite passes through the centre of the KH vortices to the left are shown in Fig. 2.

fortuitous direct encounter with the rolled-up vortices (Fig. 2) on 20 November 2001 when the upstream solar-wind magnetic field observed by the ACE spacecraft pointed northward, that is, when reconnection was less efficient but the condition was favourable for the KHI²² at the low-latitude magnetopause. The solar-wind and magnetospheric magnetic fields on the dusk flank magnetopause were approximately parallel throughout the 16 min interval (Fig. 2f, g). The period of the vortices encounter was embedded in a more than 13-h interval of quasi-periodic plasma and magnetic field perturbations related to deformations of the magnetopause (09:30–23:30 UT). Rolled-up vortices were identified far more clearly by taking full advantage of multi-spacecraft information (Fig. 2e). The flow vectors, transformed into the frame of the vortices and viewed from the north, rotate anticlockwise around the centre of the vortices (marked by the red vertical lines in Fig. 2), as expected at the duskside magnetopause. Near these vortices, the density observed by the Cluster 1 spacecraft (C1) located farthest from the magnetopause (marked by the blue arrows in Fig. 2b) toward the magnetosphere was often higher than that observed by C3 or C4 (Fig. 2c, d). These instances of higher density at C1 are marked by red bars in Fig. 2c. Such high density regions appear to be connected to the dense solar wind on the anti-sunward side, for example, at 20:27:30 UT in Fig. 2d, rather than detached from the solar-wind region. This feature is consistent with the simulation result (Fig. 1b), and indicates that the dense regions result from the

roll-up of the solar-wind plasma associated with the growth of the KHI (Fig. 1b), not from the impulsive penetration process²³. The density variation observed by C1 (Fig. 2c) is similar to that expected when a synthetic spacecraft moves through the central portion of the simulated vortex and crosses the magnetopause back and forth (Fig. 1b), suggesting that C1 was in the vicinity of the centre of the vortices.

In addition to the density and flow signatures of the rolled-up KH vortices, we identified a unique magnetic field perturbation pattern associated with these vortices, which should appear only in a 3D configuration of the magnetosphere where the KH-unstable plasma sheet is sandwiched between the KH-stable northern and southern lobes (Fig. 1a). A numerical simulation of the KHI that considers this 3D magnetosphere-like geometry predicts that this field perturbation manifests in boundary regions between the plasma sheet and the lobes. This is because only low-latitude portions of the field lines surrounding the magnetopause are engulfed into the vortices, while those at high latitudes are unaffected (see Fig. 1 legend). Figure 2f shows that magnetic field perturbations seen in the high density (solar-wind) region ($\Delta B_x > 0$ and $\Delta B_y < 0$) and in the low density (magnetospheric) region ($\Delta B_x < 0$ and $\Delta B_y > 0$) have polarities in precise agreement with the 3D KHI effect on the magnetic field on the southward side of the equatorial plane where Cluster resided (Fig. 1a). The combined plasma and magnetic field observations provide unambiguous identification of rolled-up

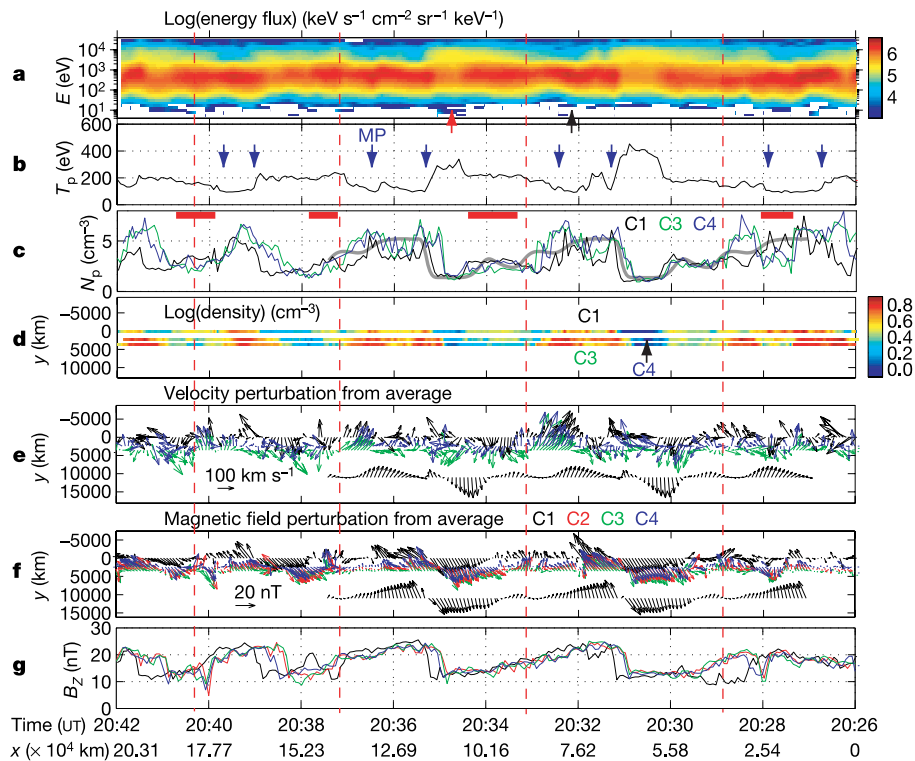


Figure 2 Detection by Cluster of rolled-up plasma vortices on 20 November 2001 (20:26–20:42 UT). **a**, The omni-directional energy spectrogram of ions observed by the Cluster 1 spacecraft (C1). Time progresses to the left, and is translated into the position of the spacecraft as follows. In Earth's rest frame, the spacecraft motion is neglected as compared to that of the vortices. $-x$ is in the direction of the vortex motion in the spacecraft frame. The vortex velocity, $\mathbf{V}_{\text{mean}}$, is computed by averaging over the above interval the ion bulk velocity vectors measured by C1, C3 and C4. The spacecraft position is determined using an equation, $x = |\mathbf{V}_{\text{mean}}|t + \Delta x_{i1}$, where t is time elapsed from the start of the interval, and Δx_{i1} is the x position of the i th spacecraft relative to C1. The arrows at the bottom denote the moments when the spectra and velocity distributions shown in Fig. 3 were observed. **b**, The ion temperature obtained by C1. The blue arrows

mark approximate locations of the magnetopause. **c**, The plasma density variations are similar to those predicted by the numerical simulation (thick grey curve). Red bars indicate instances when C1 observed higher density than C3 and C4. **d**, The plasma density colour-coded and projected along the spacecraft trajectories. y is orthogonal to both x and the direction of the averaged magnetic field, $\mathbf{B}_{\text{mean}}$, and points outward along the magnetopause normal. **e**, **f**, The x – y projection of the velocity and magnetic field deviations from $\mathbf{V}_{\text{mean}}$ and $\mathbf{B}_{\text{mean}}$, respectively (C1, black; C2, red; C3, green; and C4, blue), along with the behaviour predicted by the simulation (below). The red dashed vertical lines mark the approximate centres of the vortices. The spacecraft separation distance and the length of the arrows are doubled in the y direction. **g**, The z component of the measured magnetic field.

vortices at the magnetopause, which, according to theory, is a key (and necessary) ingredient for plasma transport via KHI.

Evidence for plasma transport across the magnetopause was indeed observed. Cold solar-wind and hot magnetospheric ion populations are found to coexist in the vortices (Figs 2a, 3). Significant amounts of the solar-wind (<2 keV) and magnetospheric (>5 keV) ions were detected simultaneously in the same region on the magnetosphere side of the magnetopause. The appearance of rolled-up vortices in the vicinity of the boundary layer is strongly suggestive of the KHI mechanism for plasma transport across the boundary.

We could not deduce the exact microphysical process that leads to plasma transport within the KH vortices on the basis of the present measurements. However, we could rule out local reconnection, because during the interval of the vortex observations, we did not find signatures of plasma acceleration due to magnetic stresses^{24,25} (see Supplementary Fig. 1) and D-shaped ion distributions characteristic of reconnection²⁶. To exclude, conclusively, the possibility of remote (high-latitude) reconnection supplying the plasma observed at low latitude is more difficult. We do note, however, that the boundary layer ions, detected off the equator, were flowing poleward in precisely the same direction as the solar-wind flow. This seems inconsistent with the idea that high-latitude magnetopause reconnection²⁷, which would result in an equatorward flow at the observation point, produced the observed boundary layer. These observations thus indicate that reconnection occurred neither locally nor at the high-latitude magnetopause during the time of the observations of KH vortices. However, we cannot rule out the possibility that reconnection occurred in the past (but had ceased to exist) to produce a boundary layer which is now rolled-up by the KHI. Proving the existence of such a scenario is, however, even more difficult.

We estimate the speed of motion of the vortices by averaging over the vortices' interval the velocity values measured by the three spacecraft (C1, C3 and C4) on which the plasma instruments were operative. Time is then translated into distance of the spacecraft from a certain position in a vortex, and the length scale of one vortex is estimated to be 40,000–55,000 km (see Fig. 2). Consequently, the

initial thickness of the velocity shear layer is inferred to have been roughly 5,000–7,000 km, because the wavelength of the fastest growing KH mode is approximately eight times the initial total thickness of the velocity shear layer²⁸. According to numerical simulations, the width of the sufficiently developed KH vortex equivalent to that of the plasma boundary layer reaches about four times the initial thickness. We therefore infer that the boundary layer with the thickness of 20,000–28,000 km had been formed in the most KH-unstable low-latitude regions near, or at least further downstream of, the observation site.

The present results indicate that the KHI occurs at the flank magnetopause and that it may lead to solar-wind entry, perhaps via non-reconnection-associated processes, for northward solar-wind magnetic field. But the microphysical process that causes the plasma transport in the KH vortices, which would control the rate at which mass and energy of the solar wind are transferred, remain to be understood. The identification of the transport processes requires high-resolution measurements capable of resolving small-scale structures and dynamics embedded in the vortices. Such observations (which will be carried out by the future NASA Magnetospheric Multiscale mission²⁹ and the Japanese SCOPE mission) would enable us to determine the relative contributions of reconnection and the KHI during northward solar-wind magnetic field conditions. □

Received 12 April; accepted 29 June 2004; doi:10.1038/nature02799.

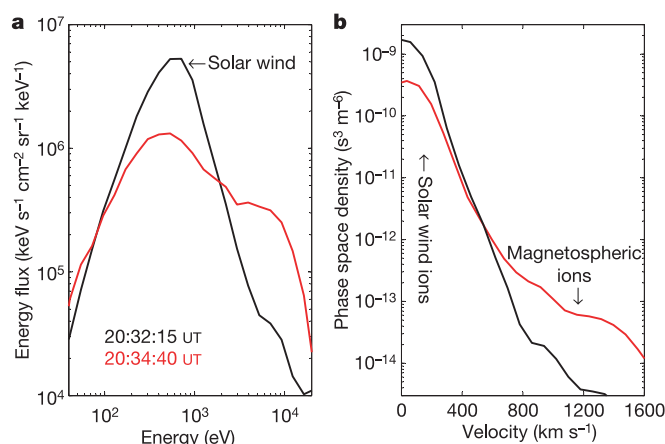


Figure 3 The ion energy spectra and velocity distributions observed by C1, showing the coexistence of the solar-wind (cold) and magnetospheric (hot) populations. The red curves were obtained at 20:34:40 UT when the spacecraft was in the magnetosphere near a vortex, while the black curves were obtained at 20:32:15 UT when the spacecraft was in the solar-wind-like region. **a**, The energy flux plotted as a function of energy in the spacecraft frame. The two peaks seen at 20:34:40 UT indicate the coexistence of two distinct populations. **b**, The velocity distributions at 90° to the magnetic field in the frame of bulk flow. The coexistence of solar-wind and magnetospheric ions in the vicinity of the vortices is suggestive of the KHI as a means of plasma transport across the magnetopause.

1. Akasofu, S.-I. Energy coupling between the solar wind and the magnetosphere. *Space Sci. Rev.* **28**, 121–190 (1981).
2. Cowley, S. W. H. in *Magnetic Reconnection in Space and Laboratory Plasmas* (ed. Hones, E. W.) 375–378 (Geophys. Monograph 30, American Geophysical Union, Washington DC, 1984).
3. Mitchell, D. G. et al. An extended study of the low-latitude boundary layer on the dawn and dusk flanks of the magnetosphere. *J. Geophys. Res.* **92**, 7394–7404 (1987).
4. Hasegawa, H., Fujimoto, M., Saito, Y. & Mukai, T. Dense and stagnant ions in the low-latitude boundary region under northward interplanetary magnetic field. *Geophys. Res. Lett.* **31**, L06802 (2004).
5. Chandrasekhar, S. *Hydrodynamic and Hydromagnetic Stability* (Oxford Univ. Press, New York, 1961).
6. Scoppe, N. G. et al. Structure of the low-latitude boundary layer. *J. Geophys. Res.* **86**, 2099–2110 (1981).
7. Song, P. & Russell, C. T. Model of the formation of the low-latitude boundary-layer for strongly northward interplanetary magnetic-field. *J. Geophys. Res.* **97**, 1411–1420 (1992).
8. Sibeck, D. G. et al. in *Magnetospheric Plasma Sources and Losses* Ch. 5 (ed. Hultqvist, B.) 207–283 (Space Sciences Series of ISSI 6, Kluwer Academic, Dordrecht, 1999).
9. Dungey, J. W. in *Proc. Ionosphere Conf.* 225 (Physical Society of London, 1955).
10. Miura, A. Anomalous transport by magnetohydrodynamic Kelvin-Helmholtz instabilities in the solar wind-magnetosphere interaction. *J. Geophys. Res.* **89**, 801–818 (1984).
11. Thomas, V. A. & Winske, D. Kinetic simulations of the Kelvin-Helmholtz instability at the magnetopause. *J. Geophys. Res.* **98**, 11425–11438 (1993).
12. Fujimoto, M. & Terasawa, T. Anomalous ion mixing within an MHD scale Kelvin-Helmholtz vortex. *J. Geophys. Res.* **99**, 8601–8613 (1994).
13. Huba, J. D. The Kelvin-Helmholtz instability: Finite Larmor radius magnetohydrodynamics. *Geophys. Res. Lett.* **23**, 2907–2910 (1996).
14. Nykyri, K. & Otto, A. Plasma transport at the magnetospheric boundary due to reconnection in Kelvin-Helmholtz vortices. *Geophys. Res. Lett.* **28**, 3565–3568 (2001).
15. Matsumoto, Y. & Hoshino, M. Onset of turbulence induced by a Kelvin-Helmholtz vortex. *Geophys. Res. Lett.* **31**, L02807 (2004).
16. Nakamura, T. K. M., Hayashi, D., Fujimoto, M. & Shinohara, I. Decay of MHD-scale Kelvin-Helmholtz vortices mediated by parasitic electron dynamics. *Phys. Rev. Lett.* **92**, 145001 (2004).
17. Ogilvie, K. W. & Fitzenreiter, R. J. The Kelvin-Helmholtz instability at the magnetopause and inner boundary layer surface. *J. Geophys. Res.* **94**, 15113–15123 (1989).
18. Kivelson, G. K. & Chen, S.-H. in *Physics of the Magnetopause* (eds Song, P., Sonnerup, B. U. Ö. & Thomsen, M. F.) 257–268 (Geophys. Monograph 90, American Geophysical Union, Washington DC, 1995).
19. Hones, E. W. Jr et al. Further determination of the characteristics of magnetospheric plasma vortices. *J. Geophys. Res.* **86**, 814–820 (1981).
20. Fairfield, D. H. et al. Geotail observations of the Kelvin-Helmholtz instability at the equatorial magnetotail boundary for parallel northward fields. *J. Geophys. Res.* **105**, 21159–21173 (2000).
21. Fujimoto, M., Tonooka, T. & Mukai, T. in *Earth's Low-latitude Boundary Layer* (eds Newell, P. T. & Onsager, T.) 241–251 (Geophys. Monograph 133, American Geophysical Union, Washington DC, 2003).
22. Miura, A. Dependence of the magnetopause Kelvin-Helmholtz instability on the orientation of the magnetosheath magnetic field. *Geophys. Res. Lett.* **22**, 2993–2996 (1995).
23. Lemaire, J. Impulsive penetration of filamentary plasma elements into magnetospheres of Earth and Jupiter. *Planet. Space Sci.* **25**, 887–890 (1977).
24. Paschmann, G. et al. Plasma acceleration at the Earth's magnetopause: Evidence for reconnection. *Nature* **282**, 243–246 (1979).
25. Phan, T.-D. et al. Extended magnetic reconnection at the Earth's magnetopause from detection of bi-directional jets. *Nature* **404**, 848–850 (2000).
26. Fuselier, S. A. in *Physics of the Magnetopause* (eds Song, P., Sonnerup, B. U. Ö. & Thomsen, M. F.) 181–187 (Geophys. Monograph 90, American Geophysical Union, Washington DC, 1995).

27. Kessel, R. L. *et al.* Evidence of high-latitude reconnection during northward IMF: Hawkeye observations. *Geophys. Res. Lett.* **23**, 583–586 (1996).
28. Miura, A. & Pritchett, P. L. Nonlocal stability analysis of the MHD Kelvin-Helmholtz instability in a compressible plasma. *J. Geophys. Res.* **87**, 7431–7444 (1982).
29. Report of the NASA Science and Technology Definition Team for the Magnetospheric Multiscale (MMS) Mission (NASA/TM-2000-209883, Goddard Space Flight Center, Greenbelt, MD, 1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are indebted to the Cluster team for the design and successful operation of the Cluster II mission. Part of this work was done while H.H. visited UC Berkeley.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.H. (hiroshi.hasegawa@dartmouth.edu).

Room-temperature ferroelectricity in strained SrTiO₃

J. H. Haeni¹, P. Irvin², W. Chang³, R. Uecker⁴, P. Reiche⁴, Y. L. Li¹, S. Choudhury¹, W. Tian⁵, M. E. Hawley⁶, B. Craigo⁷, A. K. Tagantsev⁸, X. Q. Pan⁵, S. K. Streiffer⁹, L. Q. Chen¹, S. W. Kirchoefer³, J. Levy² & D. G. Schlom¹

¹Department of Materials Science and Engineering, Penn State University, University Park, Pennsylvania 16802-5005, USA

²Department of Physics and Astronomy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260, USA

³Naval Research Laboratory, 4555 Overlook Avenue S.W., Washington DC 20375, USA

⁴Institute of Crystal Growth, Max-Born-Straße 2, D-12489 Berlin, Germany

⁵Department of Materials Science & Engineering, University of Michigan, Ann Arbor, Michigan 48109-2136, USA

⁶Materials Science and Technology Division (MST-8), Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

⁷Motorola Labs, 2100 East Elliot Road, Tempe, Arizona 85284, USA

⁸Laboratoire de Céramique, Ecole Polytechnique Fédérale de Lausanne, Lausanne CH 1015, Switzerland

⁹Materials Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA

Systems with a ferroelectric to paraelectric transition in the vicinity of room temperature are useful for devices. Adjusting the ferroelectric transition temperature (T_c) is traditionally accomplished by chemical substitution—as in $\text{Ba}_x\text{Sr}_{1-x}\text{TiO}_3$, the material widely investigated for microwave devices in which the dielectric constant (ϵ_r) at GHz frequencies is tuned by applying a quasi-static electric field^{1,2}. Heterogeneity associated with chemical substitution in such films, however, can broaden this phase transition by hundreds of degrees³, which is detrimental to tunability and microwave device performance. An alternative way to adjust T_c in ferroelectric films is strain^{4–8}. Here we show that epitaxial strain from a newly developed substrate can be harnessed to increase T_c by hundreds of degrees and produce room-temperature ferroelectricity in strontium titanate, a material that is not normally ferroelectric at any temperature. This strain-induced enhancement in T_c is the largest ever reported. Spatially resolved images of the local polarization state reveal a uniformity that far exceeds films tailored by chemical substitution. The high ϵ_r at room temperature in these films (nearly 7,000 at 10 GHz) and its sharp dependence on electric field are promising for device applications^{1,2}.

Enormous strains can be imparted to thin films and have previously been used to alter the T_c of ferromagnetic^{9,10} and super-

conducting^{11–13} materials. For such phenomena, strain-induced enhancements in T_c as large as tens of degrees have been observed⁹. Owing to the strong coupling between strain and ferroelectricity, much larger T_c shifts are expected^{4,6}, and have been observed^{7,8}, in ferroelectric materials.

In its pure, unstressed form, strontium titanate (SrTiO_3) is an incipient ferroelectric. It remains paraelectric down to 0 K, although chemical^{14,15} or isotopic substitution¹⁶, as well as the application of stress⁴, easily disturb this delicate state, resulting in ferroelectricity. The boundary conditions imposed by a substrate profoundly affect ferroelectricity in thin films. Figure 1 shows the predicted⁴ shift in T_c for SrTiO_3 (ref. 17) under biaxial strain $\epsilon_s = (a_{\parallel} - a_0)/a_0$, where a_0 is the lattice parameter of free-standing SrTiO_3 and $a_{\parallel}$ is the in-plane lattice parameter of a biaxially strained (100) SrTiO_3 film. The hatched region shows the range in predicted T_c due to the spread in reported property coefficients for SrTiO_3 (refs 18, 19) that enter into the thermodynamic analysis. For example, for positive ϵ_s and $T > 120$ K, the enhancement in T_c is given by $\Delta T_c = 2\epsilon_s\epsilon_0 C(Q_{11} + Q_{12})/(s_{11} + s_{12})$, where ϵ_0 is the permittivity of free space, C is the Curie constant, Q_{11} and Q_{12} are the electrostrictive coefficients, and s_{11} and s_{12} are elastic compliances of SrTiO_3 . The breadth of the hatched region in Fig. 1 for T_c is mainly due to the nearly factor-of-two variation in the ratio of $(Q_{11} + Q_{12})/(s_{11} + s_{12})$ for what are considered the most accurate reported values of these constants. These predictions imply that a biaxial tensile strain of order 1% will shift the T_c of SrTiO_3 to the vicinity of room temperature.

In practice, the synthesis of uniformly strained ferroelectric films is challenging. Epitaxial ferroelectric films are usually grown to thicknesses greatly exceeding their critical values, resulting in undesirable relaxation towards a zero-strain state by the introduction of dislocations. Dislocation densities of $\sim 10^{11} \text{ cm}^{-2}$ are typical in epitaxial $\text{Ba}_x\text{Sr}_{1-x}\text{TiO}_3$ films²⁰, and the resulting inhomogeneous strain further smears the phase transition, in addition to the effects of chemical heterogeneity mentioned above. Our approach to controlling the properties of ferroelectric SrTiO_3 films centres on the development of new substrates that enable the growth of uniformly strained films below, or at least far closer to, the critical thickness for relaxation.

Depending on the choice of substrate, films may be grown under compressive or tensile strain. Commercially available substrates that

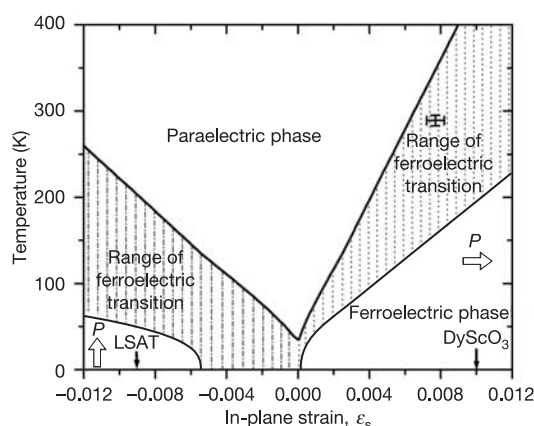


Figure 1 Expected shift in T_c of (100) SrTiO_3 with biaxial in-plane strain, based on thermodynamic analysis. The arrows indicate the predicted direction of the polarization for strained SrTiO_3 : in-plane for biaxial tensile strain and out-of-plane for biaxial compressive strain. The ϵ_s values for SrTiO_3 fully constrained (commensurate) to the lattice constants of LSAT and (110) DyScO_3 substrates are indicated by the positions of the corresponding arrows. The cross shows the observed T_c shift of a 500-Å-thick SrTiO_3 film epitaxially grown on (110) DyScO_3 .

are larger than SrTiO₃, such as KTaO₃ ($a = 3.989 \text{ \AA}$)²¹ or MgO ($4.212 \text{ \AA}$)²¹, are unsuitable for high-quality SrTiO₃ ($a = 3.905 \text{ \AA}$)¹⁸ growth because of their excessively large lattice mismatches (2.2% and 7.7%, respectively). On such substrates dislocation introduction commences soon after film nucleation, and films typically have very large mosaic spread. As the critical thickness at which dislocations begin to form varies approximately inversely with lattice mismatch, lower mismatch is desired to allow strained, low-dislocation density SrTiO₃ films to be grown that are thick enough to allow their ferroelectric properties to be conveniently probed. To produce a closer lattice match, we have developed single crystals of a new substrate material, DyScO₃, for the growth of SrTiO₃ films under uniform biaxial tensile strain. DyScO₃ is orthorhombic with lattice constants²² $a = 5.440 \text{ \AA}$, $b = 5.713 \text{ \AA}$,

and $c = 7.887 \text{ \AA}$. It is not ferroelectric down to 4 K. The (110) DyScO₃ surface has a nearly square surface mesh with an in-plane lattice spacing $a_{\parallel} = 3.944 \text{ \AA}$, resulting in a tensile lattice mismatch at room temperature (25 °C) of +1.0% with (100) SrTiO₃. For comparison, the commercial substrate (LaAlO₃)_{0.29} × (SrAl_{0.5}Ta_{0.5}O₃)_{0.71} (LSAT²³, $a = 3.869 \text{ \AA}$) with a lattice mismatch of −0.9% at room temperature was also used to grow epitaxial (100) SrTiO₃ films under uniform compressive strain. The low dielectric constants of both DyScO₃ ($\epsilon_{11} = 22.0$, $\epsilon_{22} = 18.8$ and $\epsilon_{33} = 35.5$) and LSAT ($\epsilon_r = 22.5$)²⁴ reduce field penetration in the substrate and facilitate modelling of the in-plane ϵ_r of the SrTiO₃ films when interdigitated electrodes are used²⁵, as in this study.

Reactive molecular beam epitaxy (MBE) was used to grow epitaxial SrTiO₃ films on both (110) DyScO₃ and (100) LSAT substrates. During growth the substrates were held at 650 °C and immersed in 5.0×10^{-7} torr (background pressure) of oxygen plus ~10% ozone. Reflection high-energy electron diffraction (RHEED) intensity oscillations were monitored to ensure that complete SrO and TiO₂ monolayers were deposited alternately to form the SrTiO₃ films²⁶. Although many samples were grown and characterized, the properties of a 500-Å-thick film grown on each substrate are compared below to show the effect of strain state on the ferroelectric properties of SrTiO₃.

X-ray diffraction scans show the out-of-plane lattice constant to be compressed ($a_{\perp} = 3.88 \pm 0.01 \text{ \AA}$) for the films grown on DyScO₃, and extended ($a_{\perp} = 3.93 \pm 0.01 \text{ \AA}$) for the films grown on LSAT, as expected from the boundary conditions imparted by the substrates. High-resolution X-ray diffraction measurements of the 500-Å-thick SrTiO₃ film grown on DyScO₃ reveal that its in-plane lattice constant at room temperature is $a_{\parallel} = 3.935 \pm 0.002 \text{ \AA}$, corresponding to 0.8% biaxial tensile strain. Its out-of-plane lattice constant is $a_{\perp} = 3.8785 \pm 0.0005 \text{ \AA}$. Although somewhat beyond the critical thickness at which relaxation begins to occur, the rocking curve full-width at half-maximum (FWHM) of the 200 SrTiO₃ X-ray diffraction peak is 13 arcsec (0.004°), the narrowest ever reported for an epitaxial SrTiO₃ film. This, and other aspects of the X-ray scattering measurements, indicate that this film has a higher degree of uniformity and structural perfection than is typical of Ba_xSr_{1-x}TiO₃ films. The rocking curve FWHM of the 200 SrTiO₃ peak of the 500-Å-thick SrTiO₃ film grown on LSAT was also quite narrow, 21.5 arcsec (0.006°).

Measurements of the complex dielectric permittivity were performed on the SrTiO₃ films as a function of frequency (1–20 GHz), temperature (75–330 K), and direct current (d.c.) electric field

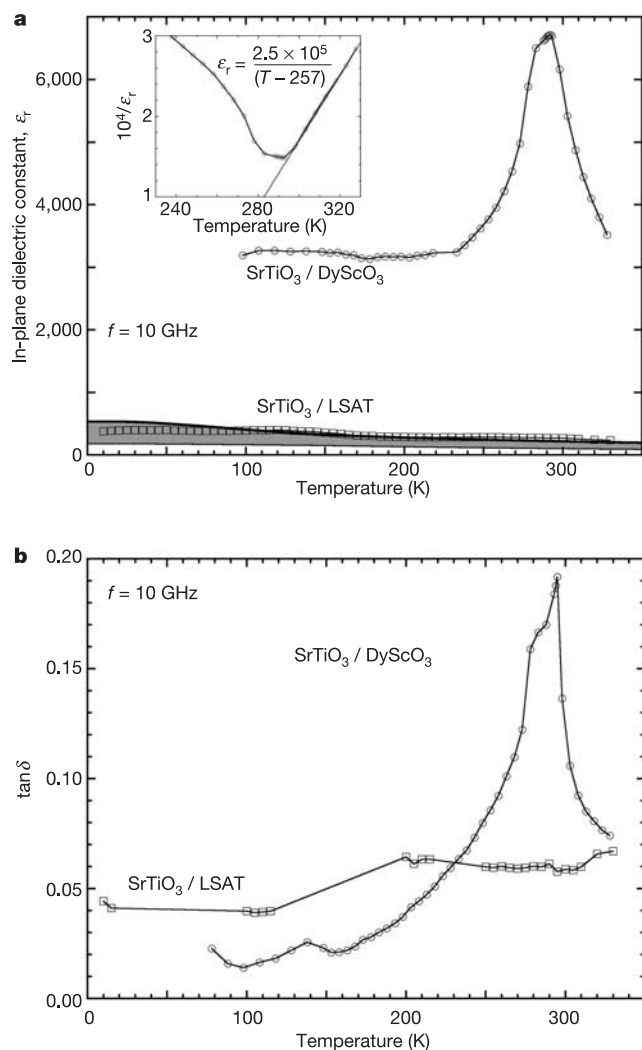


Figure 2 In-plane dielectric constant (ϵ_r) and dielectric loss ($\tan\delta$) in strained epitaxial SrTiO₃ films as a function of temperature and film thickness at a measurement frequency (f) of 10 GHz. **a**, **b**, Contrast of in-plane ϵ_r and $\tan\delta$, respectively, of 500-Å-thick SrTiO₃/(110) DyScO₃ and SrTiO₃/(100) LSAT epitaxial films. These films are under biaxial tensile and compressive strain, respectively. The peak in ϵ_r of about 7,000 and the simultaneous peak in $\tan\delta$ indicate that the T_c of SrTiO₃ under biaxial tension of $\epsilon_s = 0.008$ is about 293 K. The inset in **a** shows a Curie–Weiss fit to $1/\epsilon_r$. Owing to systemic errors involved in the measurement and calculation of the in-plane dielectric constant, the vertical scale in **a** may be off by as much as 10%. The shaded region in **a** corresponds to the expected value of the in-plane ϵ_r for a SrTiO₃ film commensurately strained to LSAT ($\epsilon_s = -0.009$), based on thermodynamic analysis and the range of relevant reported property coefficients for SrTiO₃ (refs 18, 19).

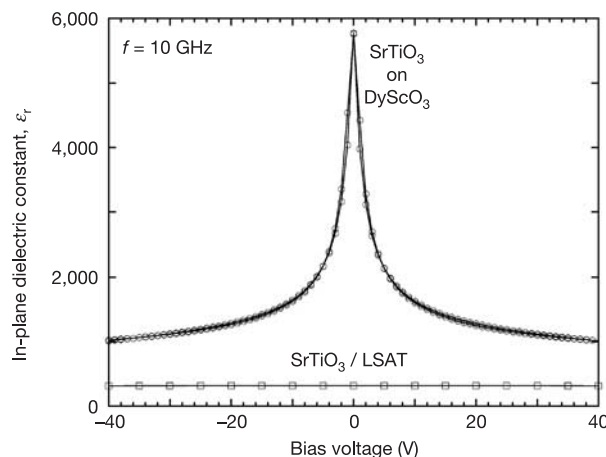


Figure 3 Dielectric tunability of the same 500-Å-thick SrTiO₃ films grown on DyScO₃ and LSAT as shown in Fig. 2. The film grown on DyScO₃ shows 82% tuning between 0 and ± 40 V at room temperature, with slight hysteresis. The film grown on LSAT exhibits ~0% tunability with the application of ± 40 V at room temperature.

(± 40 V across $6\text{ }\mu\text{m}$) using interdigitated electrodes. The 1–2- μm -thick Ag interdigitated electrodes had 80- μm finger lengths and 6- μm finger gaps for the room-temperature measurements and 40- μm finger lengths and 12- μm finger gaps for the temperature-dependent measurements. The fingers were aligned parallel to the $\langle 110 \rangle$ in-plane directions of the SrTiO_3 films.

Figure 2 shows the temperature-dependent in-plane ϵ_r and loss tangent, $\tan\delta$, of the 500-Å $\text{SrTiO}_3/\text{DyScO}_3$ film measured at 10 GHz. These data are from a device acting as a microwave varactor (a voltage-tunable capacitor) with 2:1 tuning, demonstrating the technological significance of these films. The curves agree well with expected Curie–Weiss behaviour (inset) for a uniformly strained film, exhibiting a maximum in ϵ_r and $\tan\delta$ at $T_c \approx 293\text{ K}$ and a Curie constant typical of displacive ferroelectrics. The peak ϵ_r of nearly 7,000 at 10 GHz is higher and the width of the peak in ϵ_r versus T is narrower than any value reported for $\text{Ba}_{1-x}\text{Sr}_x\text{TiO}_3$ thin films. By contrast, $\text{SrTiO}_3/\text{LSAT}$ has a much lower dielectric constant and there is no peak in ϵ_r versus T . This result is fully consistent with the biaxial strain state causing the ferroelectric polarization to lie in the plane of the film for SrTiO_3 on DyScO_3 , the direction sensed by the interdigitated electrodes, but lying out-of-plane of the SrTiO_3 grown on LSAT.

The observed T_c shift surpasses all prior records^{7,8} and gives credence to prior work on $\text{SrTiO}_3/\text{SrZrO}_3$ superlattices where ferroelectricity at room temperature in 8-nm-thick strained SrTiO_3 layers was assumed to be responsible for the dielectric properties of the composite superlattice²⁷. Our results are consistent with a strain-induced phenomenon; the possibility that minute Sc doping from the DyScO_3 substrate is in part responsible for the dramatic changes in dielectric behaviour is also being investigated, although no such effects have ever been reported²⁸.

The remarkable field sensitivity of ϵ_r , which drops by 82% for an

applied voltage of 40 V across $6\text{ }\mu\text{m}$ gaps, is shown in Fig. 3. Such large tuning with modest fields makes these films applicable to a variety of microwave applications. As is typical of ferroelectric thin films, their properties depend on thickness (Supplementary Information).

Optical measurements of the ferroelectric polarization at microwave driving fields were performed using time-resolved confocal scanning optical microscopy (TRCSOM)^{29,30}. The in-phase and out-of-phase response at the microwave driving frequency correspond to the real and imaginary parts of the polar contribution to the dielectric response at that frequency.

Figure 4c shows a typical TRCSOM image of the phase shift between linear electro-optic response and a microwave driving field (3.27 GHz), measured for the 500 Å $\text{SrTiO}_3/\text{DyScO}_3$ film. The electro-optic phase is computed from a fit to the temporal response at various points in the microwave drive cycle^{29,30}. Most striking are the large regions in which the phase is essentially uniform, separated by sharp boundaries at which the phase changes by 90° or 180° . These boundaries correspond to features observed by atomic force microscopy (AFM) on the film (Fig. 4a) as well as on bare DyScO_3 substrates, and may be due to surface imperfections that break the in-plane symmetry and produce an easy axis for the SrTiO_3 films. For sake of comparison, Fig. 4d shows a typical TRCSOM image (at a fixed time delay) for a ferroelectric film ($\text{Ba}_{0.5}\text{Sr}_{0.5}\text{TiO}_3/\text{MgO}$) that has not been strain-engineered. The strong fluctuations in the electro-optic response on submicrometre scales are typical of films that are inhomogeneously strained.

The remarkable dielectric properties of SrTiO_3 , while recognized for many decades, have previously only been accessible in the bulk and at cryogenic temperatures. Efforts to shape the properties of ferroelectric films have been restricted by available substrates. Through controlled substrate engineering, often overlooked in the growth of thin films, we have achieved ϵ_r and tunability at room temperature in SrTiO_3 films with properties comparable to bulk SrTiO_3 at cryogenic temperatures. In addition to radically enhanced dielectric properties, we expect other ferroelectric properties (including pyroelectric, piezoelectric and nonlinear optical) of SrTiO_3 to be accessible at room temperature through the application of appropriate homogeneous strain. □

Received 12 March; accepted 22 June 2004; doi:10.1038/nature02773.

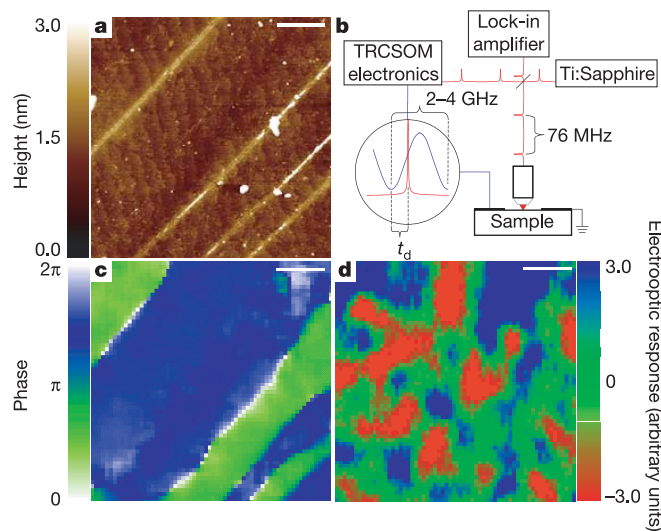


Figure 4 Comparison of the morphology and microwave electro-optic response of tunable dielectric films at room temperature (just above T_c). The scale bars are $1\text{ }\mu\text{m}$ for all images. **a**, An AFM image of the same 500-Å-thick SrTiO_3 film grown on DyScO_3 shown in Fig. 2. In addition to unit-cell-high steps on the surface of the SrTiO_3 film, four diagonal linear features that protrude $\sim 50\text{ Å}$ from the surface are evident. Similar surface imperfections are seen on the bare substrates, suggesting that they are associated with polishing. **b**, Schematic of the TRCSOM technique used for imaging ferroelectric polarization at microwave frequencies. **c**, Images of the phase shift between the electro-optic response and microwave driving field (3.27 GHz) for the same $\text{SrTiO}_3/\text{DyScO}_3$ film (but not the identical region) as in **a**, revealing uniform in-plane polarization separated by clear boundaries like the ones observed by AFM. **d**, TRCSOM image (at one time delay) of a typical epitaxial $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{TiO}_3/\text{MgO}$ film at 2.06 GHz, showing inhomogeneous in-plane polarization.

- Vendik, O. G. *Ferroelectrics for Microwave Applications* (Soviet Radio, Moscow, 1979).
- Lancaster, M. J., Powell, J. & Porch, A. Thin-film ferroelectric microwave devices. *Supercond. Sci. Technol.* **11**, 1323–1334 (1998).
- Hubert, C. *et al.* Confocal scanning optical microscopy of $\text{Ba}_{0.5}\text{Sr}_{0.5}\text{TiO}_3$ thin films. *Appl. Phys. Lett.* **71**, 3353–3355 (1997).
- Devonshire, A. F. Theory of ferroelectrics. *Phil. Mag. Suppl.* **3**, 85–130 (1954).
- Uwe, H. & Sakudo, T. Stress-induced ferroelectricity and soft phonon modes in SrTiO_3 . *Phys. Rev. B* **13**, 271–286 (1976).
- Pertsev, N. A., Zembilgotov, A. G. & Tagantsev, A. K. Effect of mechanical boundary conditions on phase diagrams of epitaxial ferroelectric thin films. *Phys. Rev. Lett.* **80**, 1988–1991 (1998).
- Abe, K. *et al.* Modification of ferroelectricity in heteroepitaxial (Ba,Sr) TiO_3 films for non-volatile memory applications. *Integr. Ferroelectr.* **21**, 197–206 (1998).
- Streiffer, S. K. *et al.* Observation of nanoscale 180° stripe domains in ferroelectric PbTiO_3 thin films. *Phys. Rev. Lett.* **89**, 067601 (2002).
- Beach, R. S. *et al.* Enhanced Curie temperatures and magnetoelastic domains in Dy/Lu superlattices and films. *Phys. Rev. Lett.* **70**, 3502–3505 (1993).
- Gan, Q., Rao, R. A., Eom, C. B., Garrett, J. L. & Lee, M. Direct measurement of strain effects on magnetic and electrical properties of epitaxial SrRuO_3 thin films. *Appl. Phys. Lett.* **72**, 978–980 (1998).
- Lock, J. M. Penetration of magnetic fields into superconductors. III. Measurements on thin films of tin, lead and indium. *Proc. R. Soc. Lond. A* **208**, 391–408 (1951).
- Sato, H. & Naito, M. Increase in the superconducting transition temperature by anisotropic strain effect in (001) $\text{La}_{1.85}\text{Sr}_{0.15}\text{CuO}_4$ thin films on LaSrAlO_4 substrates. *Physica C* **274**, 221–226 (1997).
- Bozovic, I., Logvenov, G., Belca, I., Narimbetov, B. & Svedko, I. Epitaxial strain and superconductivity in $\text{La}_{2-x}\text{Sr}_x\text{CuO}_4$ thin films. *Phys. Rev. Lett.* **89**, 107001 (2002).
- Bednorz, J. G. & Müller, K. A. $\text{Sr}_{1-x}\text{Ca}_x\text{TiO}_3$: An XY quantum ferroelectric with transition to randomness. *Phys. Rev. Lett.* **52**, 2289–2292 (1984).
- Ang, C. *et al.* Dielectric polarization processes in Bi:SrTiO_3 . *J. Phys. Chem. Solids* **61**, 191–196 (2000).
- Itoh, M. *et al.* Ferroelectricity induced by oxygen isotope exchange in strontium titanate perovskite. *Phys. Rev. Lett.* **82**, 3540–3543 (1999).
- Pertsev, N. A., Tagantsev, A. K. & Setter, N. Phase transitions and strain-induced ferroelectricity in SrTiO_3 epitaxial thin films. *Phys. Rev. B* **61**, R825–R829 (2000).
- Hellwege, K.-H. & Hellwege, A. M. (eds) *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology New Series, Group III, Vol. 16a*, 59–64 (Springer, Berlin, 1981).

19. Hellwege, K.-H. & Hellwege, A. M. (eds) *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology* New Series, Group III, Vol. 11, 418 (Springer, Berlin, 1979).
20. Canedy, C. L. *et al.* Dielectric properties in heteroepitaxial $\text{Ba}_{0.6}\text{Sr}_{0.4}\text{TiO}_3$ thin films: Effect of internal stresses and dislocation-type defects. *Appl. Phys. Lett.* **77**, 1695–1697 (2000).
21. Hellwege, K.-H. & Hellwege, A. M. (eds) *Landolt-Börnstein: Numerical Data and Functional Relationships in Science and Technology* New Series, Group III, Vol. 7b, 26; Vol. 7c, 635 (Springer, Berlin, 1976).
22. JCPDS Powder Diffraction File: Sets 27 to 28 72 (Card 27–204, JCPDS International Centre for Diffraction Data, Swarthmore, 1986).
23. Chakoumakos, B. C., Schlom, D. G., Urbanik, M. & Luine, J. Thermal expansion of LaAlO_3 and $(\text{La,Sr})(\text{Al,Ta})\text{O}_3$ substrate materials for superconducting thin-film device applications. *J. Appl. Phys.* **83**, 1979–1982 (1998).
24. Tidrow, S. C. *et al.* New substrates for HTSC microwave devices. *IEEE Trans. Appl. Supercond.* **7**, 1766–1768 (1997).
25. Gevorgian, S. S., Martinsson, T., Linnér, P. L. J. & Kollberg, E. L. CAD models for multilayered substrate interdigital capacitors. *IEEE Trans. Microwave Theory Tech.* **44**, 896–904 (1996).
26. Haeni, J. H., Theis, C. D. & Schlom, D. G. RHEED intensity oscillations for the stoichiometric growth of SrTiO_3 thin films by reactive molecular beam epitaxy. *J. Electroceram.* **4**, 385–391 (2000).
27. Christen, H.-M., Knauss, L. A. & Harshavardhan, K. S. Field-dependent dielectric permittivity of paraelectric superlattice structures. *Mater. Sci. Eng. B* **56**, 200–203 (1998).
28. Lemanov, V. V., Sotnikov, A. V., Smirnova, E. P. & Weihnacht, M. Giant dielectric relaxation in SrTiO_3 – $\text{SrMg}_{1/3}\text{Nb}_{2/3}\text{O}_3$ and SrTiO_3 – $\text{SrSc}_{1/2}\text{Ta}_{1/2}\text{O}_3$ solid solutions. *Fiz. Tverd. Tela (St Petersburg)* **44**, 1948–1957 (2002).
29. Hubert, C. & Levy, J. New optical probe of GHz polarization dynamics in ferroelectric thin films. *Rev. Sci. Instrum.* **70**, 3684–3687 (1999).
30. Hubert, C., Levy, J., Cukauskas, E. J. & Kirchoefer, S. W. Mesoscopic microwave dispersion in ferroelectric thin films. *Phys. Rev. Lett.* **85**, 1998–2001 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We acknowledge discussions and interactions with M. D. Biegalski, J. Schubert, S. Trolier-McKinstry and J. Mannhart during the course of this work. In addition, the financial support of the National Science Foundation, the Office of Naval Research for the work performed at NRL, the Swiss National Science Foundation, and, for the work performed at ANL, the US Department of Energy, Basic Energy Sciences—Materials Sciences is gratefully acknowledged.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.G.S. (schlom@ems.psu.edu).

Bulk glasses and ultrahard nanoceramics based on alumina and rare-earth oxides

A. Rosenfanz, M. Frey, B. Endres, T. Anderson, E. Richards & C. Schardt

3M Central Research Laboratory, Corporate Research & Development,
3M Company, St Paul, Minnesota 55144-1000, USA

Although often regarded as a network-former in conventional silicate glasses, Al_2O_3 alone cannot be obtained as a bulk glass. Until now, glasses comprising continuously linked $[\text{AlO}_x]$ polyhedra have been prepared in only a few systems under very fast cooling conditions, which limits their dimensions to a few millimetres^{1–3}. Yet it is desirable to prepare bulk, or monolithic, alumina-rich glasses, with the prospect of superior mechanical, chemical and optical properties⁴. Here we report a novel process for preparing very-high-alumina glasses and nanoscale glass-ceramics. Fully dense bulk articles in net shape are obtained through viscous sintering of glass microbeads. Additional heat treatment of the consolidated glasses leads to fully crystallized transparent glass-converted nanoceramics with a hardness similar to that of alumina. This method avoids the impracticably high applied pressures (more than 1 GPa) that have been required in most cases to prepare nanocrystalline ceramics by sintering, owing to the concurrent nature of densification and grain growth under pressureless conditions^{5,6}. The reported techniques can be

extended to form glasses and nanoceramics in other oxide systems that do not include a conventional glass-forming component.

Elemental and oxide melts lacking conventional glass-formers such as sulphur or silica are prone to crystallization during slow cooling. To form glasses they must be 'kinetically frozen' by rapid quenching, as originally shown for metals⁷. Typically, slower cooling rates can be used for multicomponent eutectic compositions in which ratios of glass transition temperatures T_g relative to the melting point T_m are higher^{8,9}. For example, Al_2O_3 , a reluctant glass former, necessitates quenching rates as high as 10^7 K s^{-1} to retain an amorphous state^{10,11}. However, when alloyed in eutectic proportions with other oxides, notably CaO or RE_2O_3 (where RE stands for rare earth), glasses can be generated at cooling rates of less than 10^3 K s^{-1} (refs 3, 12, 13). These rates are nevertheless still too high to obtain bulk (more than 1 cm) glasses, limiting the achievable forms to thin films, ribbons, fibres and small spheres.

Sintering of conventional glass frit into coherent, dense, bulk forms is a well-known method in which free surfaces are eliminated by viscous flow at temperatures above T_g . For glasses that do not devitrify, there is little restriction on the sintering temperatures. However, many quenched glasses crystallize rapidly at temperatures T_x when the supplied thermal energy mobilizes the structure and the latent heat previously trapped during quenching is released. The working temperature of such glasses is thus restricted to between T_g and T_x . The successful exploitation of this kinetic window, $\Delta T_x = T_x - T_g$, was instrumental in the development of bulk metallic glasses from glassy powders a decade ago¹⁴, but it had not been applied to non-silicate oxides. However, thermal analysis data indicate that a sizable window of ΔT_x exists in a rare-earth–alumina system, and 'jelly-like' formability of Al_2O_3 – Gd_2O_3 glass fibres, when heated inside the ΔT_x regime, has been reported¹⁵. Here we show that the compositional control of this kinetic window provides an easy path to the preparation of bulk glasses and nanocrystalline glass-ceramics in a variety of non-conventional glass-forming oxide systems.

Alumina–rare-earth oxide systems (Al_2O_3 – RE_2O_3) were chosen for this study because of their technological importance. Binary

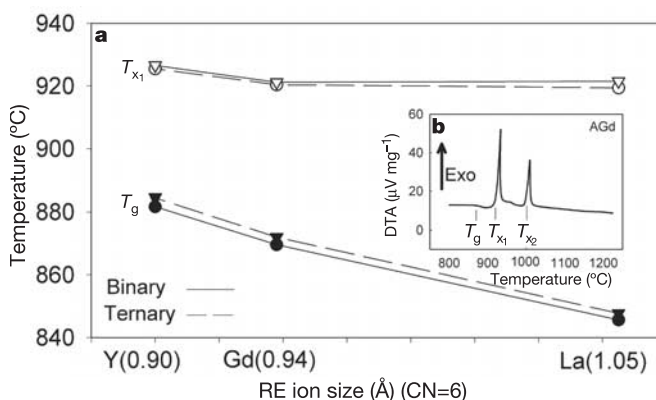


Figure 1 Effect of rare-earth ion size on ΔT_x . **a**, Glass transition (T_g) and crystallization (T_x) temperatures of binary (solid lines) and ternary (dashed lines) glasses as a function of the ionic radius of the rare-earth (RE) cation. CN, coordination number. **b**, DTA trace of 77 Al_2O_3 –23 Gd_2O_3 (mol. %) glass heated at $10^\circ \text{C min}^{-1}$. Although two exothermic events— T_{x1} and T_{x2} —were frequently observed in these glasses, only the first (T_{x1}) reflects the crystallization event, whereas T_{x2} is thought to be caused by a liquid–liquid transition from a metastable high-density amorphous phase to more stable low-density amorphous phase²⁷. In ternary liquids, the T_{x2} peak was essentially absent, which might have been caused by higher viscosities of these liquids and the retention of a single (high-density amorphous) phase during cooling. Throughout the article, T_x is identified with T_{x1} .

eutectic compositions of a high Al_2O_3 content, with RE = La, Gd and Y, were prepared with 77.5, 77 and 81.8 mol% Al_2O_3 , respectively. (The balance was RE_2O_3 .) Ternary compositions with ZrO_2 addition, at $\text{Al}_2\text{O}_3\text{:RE}_2\text{O}_3\text{:ZrO}_2$ molar ratios of 57.3:19.7:23 (La), 65.3:19.5:15.2 (Gd) and 65:16:19 (Y), were also investigated. These materials will be referred to hereafter by the first letters of their oxide components (for example, ALZ for $\text{Al}_2\text{O}_3\text{--La}_2\text{O}_3\text{--ZrO}_2$ composition). For conciseness, only the aforementioned compositions are discussed in detail here. However, similar results can be achieved for a range of previously reported $\text{Al}_2\text{O}_3\text{--RE}_2\text{O}_3$ glass compositions^{2,3}.

Glass microspheres were prepared by a flame-spraying technique in which the particulate precursors (wet-mixed, dried, crushed and sieved oxide powders) were fed into a high-temperature $\text{H}_2\text{--O}_2$ flame, producing molten particles that were quenched in water at an estimated cooling rate of 10^3 K s^{-1} . The flame-sprayed beads were less than $140 \mu\text{m}$ in diameter (see Supplementary Fig. 1). The size fraction between 75 and $109 \mu\text{m}$ was separated (by sieving) for further use. Densification of beads by pressureless sintering or hot pressing was conducted in a graphite furnace with a nitrogen atmosphere. Properties of glass beads, bulk glasses and glass-ceramics were evaluated by optical microscopy, scanning electron microscopy, X-ray diffraction (XRD) and differential thermal analysis (DTA). Hardness (H_v) and fracture toughness (K_{IC}) were measured by the Vickers indentation method under loads of 300 g for glasses and 1 kg for glass-ceramics.

More than 95% of the flame-sprayed beads were transparent and amorphous by XRD for all the binary and ternary compositions reported. The yield of transparent beads increased with increasing RE ion size (La > Gd > Y). To quantify this apparent trend of glass-forming ability further, we plot in Fig. 1a the glass transition (T_g) and crystallization (T_x) temperatures, showing a widening gap of ΔT_x from 50 to 80 K, progressing from Y to La. The observed RE trend suggests that the retention of a liquid state is linked to the size of ionic species, probably through their effect on atomic packing density. A large size difference between the constituent cations has been thought to enhance short-range ordering and promote glass formation^{16,17}, which is consistent with the progression in ionic radius for the present series, $\text{La}^{3+} > \text{Gd}^{3+} > \text{Y}^{3+} > \text{Al}^{3+}$. In contrast, a larger RE depresses T_g (see Fig. 1a) suggesting a less rigid glass network. Such a dependence of the T_g on RE size closely parallels that in the RE–aluminosilicate glasses¹⁸, even though no SiO_2 exists in our compositions. We conclude that the lower field

strength of the larger RE ion has the same weakening effect on the cross-linking networks in both silicate and aluminate glasses. Here we exploited the kinetic window between T_g and T_x to consolidate glass beads into bulk, transparent articles, such as those shown in Fig. 2a–e. After densification, the bulk glasses remain amorphous, as demonstrated by XRD and DTA measurements (Fig. 2f, g), which are nearly identical to those of the starting glass beads. The densified bodies exhibited glassy, conchoidal fracture. Fully consolidated samples of all of the glasses were produced at a deformation temperature (T_d) of 925°C and an applied uniaxial pressure of 34 MPa, as shown in Fig. 3b for three $10 \text{ mm} \times 20 \text{ mm}$ cylinders with AL, AG and AY compositions. Under the same stress, the halftimes $\tau_{1/2}$ of the densification curves were found to depend strongly on the difference between T_d (deformation temperature) and T_g ; see Fig. 3a. In viscous sintering, densification rates are inversely proportional to viscosity η according to the Frenkel equation ($d\rho/dt = 9\gamma/(4\eta r)$, where γ is the surface energy and r is the particle diameter). The correlation in Fig. 3a therefore suggests a scaling relation between η and excess temperature, $\Delta T_d = T_d - T_g$. Such behaviour is consistent with conventional oxide glasses for which the viscosity, described by the Vogel–Fischer–Tamman (VFT) equation ($\eta = \eta_0 \exp(A/(T - T_0))$, where η_0 , A and T_0 are constants), reduces to the same $\eta - \Delta T_d$ scaling relation in the vicinity of the glass transition.

Because of the lower T_g associated with the larger RE, the viscosity at the same deformation temperature decreases from Y to La as ΔT_d increases, a trend clearly seen in Fig. 3a. However, when compared at the same excess temperature, the glass with a larger RE actually exhibits a measurably longer $\tau_{1/2}$, indicating a higher viscosity (see Fig. 3a at $\Delta T_d = 55 \text{ K}$). This implies that there is a more rapid decrease in the liquid viscosity above T_g in the glasses containing a smaller RE, despite their higher field strength and T_g . Such glasses might be regarded as more ‘fragile’ because their apparent activation energy for viscous flow decreases more rapidly with temperature¹⁹. This rheological signature of fragility is probably linked to the diminished glass-forming ability and smaller ΔT_x in Fig. 1a for glasses with smaller RE.

Pressureless sintering also induced glass bead consolidation by means of viscous flow. As direct microscopic evidence of viscous sintering, Fig. 4a shows neck filling between glass beads of ALZ composition after sintering at 945°C for 10 min. The driving force for sintering is capillary pressure, $2\gamma/r^*$, where r^* is the neck radius of curvature. Assuming a surface energy of 3 J m^{-2} , this pressure is estimated to be about 0.3 MPa for the microbeads shown in Fig. 4a. Pressureless sintering becomes progressively slower as the neck

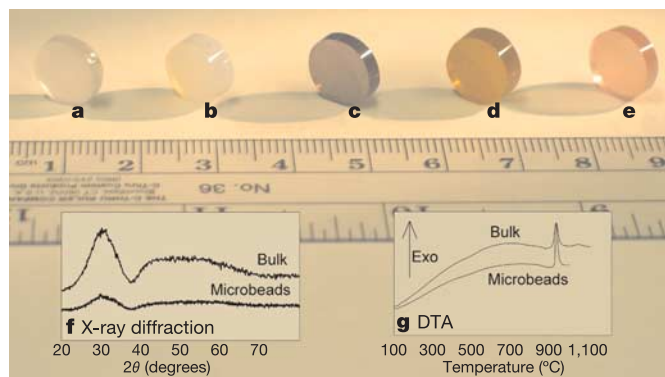


Figure 2 Bulk rare-earth aluminate glasses. **a–e**, Optical photograph of bulk ALZ materials. **a**, no dopants used; **c**, 5wt% Nd_2O_3 ; **d**, 5wt% Eu_2O_3 ; **e**, 5wt% Er_2O_3 . All except **b** were hot-pressed at 905°C at 34 MPa for 360 s. Material **b** was hot-pressed similarly but with the holding time extended to 1,200 s, inducing partial crystallization, giving the opalescent appearance. **f**, **g**, X-ray diffraction patterns (**f**) and DTA traces (**g**) of the microbead and bulk forms of **a**, showing their amorphous nature and similar thermal behaviour.

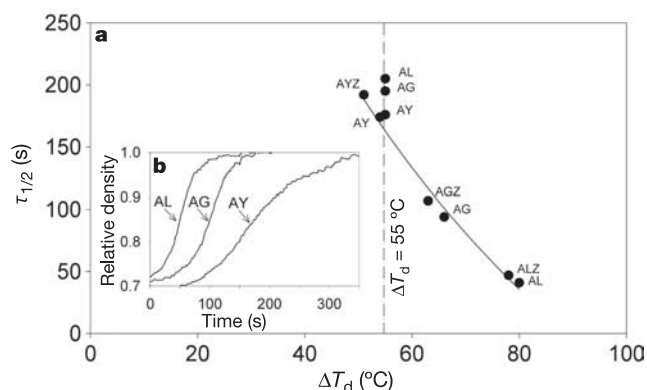


Figure 3 Consolidation of glass microbeads by hot-pressing. **a**, Plot of ΔT_d against half-time of densification, $\tau_{1/2}$. **b**, $\tau_{1/2}$ is defined as the inflection point on the relative-density–time curve indicated by the arrows for AL, AG and AY compositions, hot-pressed at 925°C at 34 MPa.

radius increases. To avoid devitrification during long-duration sintering, we chose to use a modest pressure (from 1 to 50 MPa) to aid sintering, enabling the net shape forming of intricate articles.

Conversion of the bulk glasses to nanoscale glass-ceramics was achieved by heating above T_x for a short time, resulting in simultaneous crystallization and grain growth (see Supplementary Fig. 2). For example, grain size increased from several nanometres at the early stages of conversion (900–1,000 °C) to just over 100 nm at its conclusion (1,150–1,300 °C), yielding the final microstructure shown in Fig. 4b. As in conventional glass-ceramics, the microstructure can be tailored by phase assemblage and heat treatment. For instance, by using the same ALZ composition but extending the heat treatment time at 1,300 °C from 0.5 to 3 h, the fine, equiaxed-grain microstructure (made of LaAlO_3 , transitional Al_2O_3 and t-ZrO_2) in Fig. 4b was replaced by a duplex microstructure with larger acicular $\text{LaAl}_{11}\text{O}_{18}$ crystals in a matrix of fine-grained LaAlO_3 , transitional Al_2O_3 and ZrO_2 (Fig. 4c). This microstructure transition was accompanied by a crossover in mechanical characteristics from a state of higher hardness (18.3 GPa) but lower toughness ($2.1 \text{ MPa m}^{1/2}$) to one of lower hardness (14.4 GPa) and higher toughness ($4.2 \text{ MPa m}^{1/2}$), a phenomenon typical of other glass-ceramic systems²⁰. Compared with ceramics of the same composition processed by other methods, the present glass-ceramic route provides finer and, most importantly, more homogeneous microstructures. For example, when materials with RE–aluminate eutectic compositions are directly cooled from the melt into articles with dimensions measuring several millimetres, eutectic colonies with sizes ranging from 5 to 20 μm are formed²¹. The mechanical properties of the latter materials are limited by the larger sizes of the colonies and do not fully benefit from their finer spacing of lamellae (less than 0.5 μm).

In conventional nanophase glass-ceramics²² some residual

glass regions typically remain, adversely affecting the mechanical and chemical properties of the otherwise crystallized material. Owing to kinetic limitations of crystallization, the amorphous phase also remains in Al_2O_3 –REO glasses after the initial exothermic event (T_{x1} in Fig. 1b)¹⁵. However, the absence of a strong glass-former such as silica allows complete devitrification at higher temperatures to yield fully crystalline, extremely hard nanoscale materials. To compare these glass-converted ceramics with ceramics from the conventional powder route, we prepared a fine-grain translucent alumina by isostatic hot-pressing crystalline powders²³ (see Fig. 4d). The hot-pressed alumina had an average grain size of 0.8 μm and excellent mechanical properties ($H_v = 18.8 \text{ GPa}$, $K_{Ic} = 3.4 \text{ MPa m}^{1/2}$), which represents the state of the art for corundum-based ceramics. As shown in Fig. 5, Al_2O_3 –REO glass-ceramics have comparable hardness, despite the presence of the much softer, non-corundum, REO-containing phases (REAlO_3 , $\text{REAl}_{11}\text{O}_{18}$ and $\text{RE}_3\text{Al}_5\text{O}_{12}$), which typically have a hardness of less than 14 GPa. Such a remarkable hardening effect in the alumina–REO nanocomposite is a vivid demonstration of the dominance of interfacial processes over the crystal lattice effects in the nanosize domain, as described recently²⁴.

The benefit of the glass route to the presently reported nanoscale glass-ceramics extends beyond mechanical properties. As is already evident from Fig. 2a–e, the glasses are transparent and can exhibit various colours and optical activities (for example, luminescence) as a result of RE doping (see Supplementary Figs 3 and 4). Similarly, when the grain size is maintained below the scattering limit, the fully crystallized Al_2O_3 –REO ceramics exhibit attractive optical characteristics including high refractive index (1.8 and higher) and transparency through the mid-infrared range.

Finally, the discovery of glass-forming ability and glass-converted nanoscale ceramics in the Al_2O_3 –REO and Al_2O_3 – ZrO_2 –REO systems can be extended to other non-conventional bulk oxide systems, provided a sufficiently wide kinetic window ΔT_x is available. The width of this window apparently depends on the size difference of the cations involved, as illustrated by the following examples using $\text{Ba}^{2+} > \text{La}^{3+} > \text{Ca}^{2+} > \text{Zr}^{4+} > \text{Ti}^{4+} > \text{Al}^{3+}$. In the Al_2O_3 , Al_2O_3 – TiO_2 and Al_2O_3 – ZrO_2 systems there is no

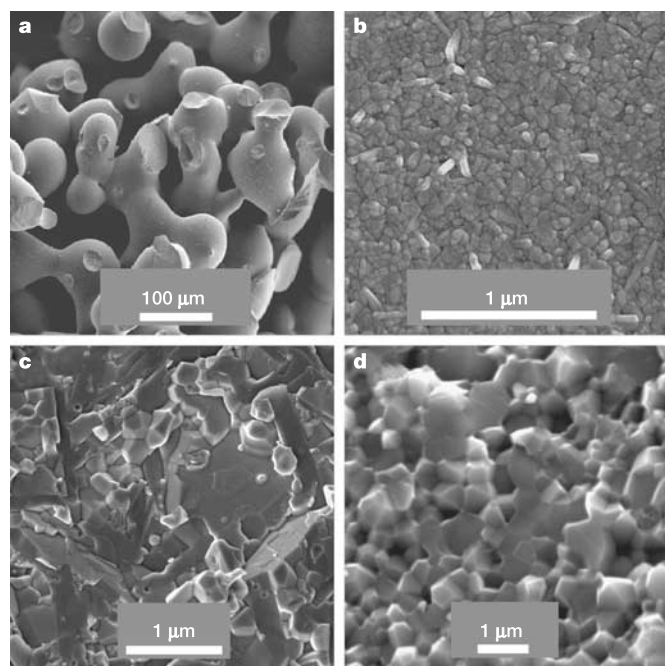


Figure 4 Scanning electron micrographs of fracture surfaces. **a**, ALZ material sintered without pressure at 945 °C for 600 s. **b**, ALZ material consolidated at 925 °C for 100 s at 34 MPa, then heated at 10 °C min^{-1} to 1,300 °C and held for 0.5 h for crystallization. **c**, The heat treatment of **b** at 1,300 °C was extended to 3 h. (Acicular grains are identified as $\text{LaAl}_{11}\text{O}_{18}$ phase.) **d**, Translucent Al_2O_3 prepared in accordance with ref. 23.

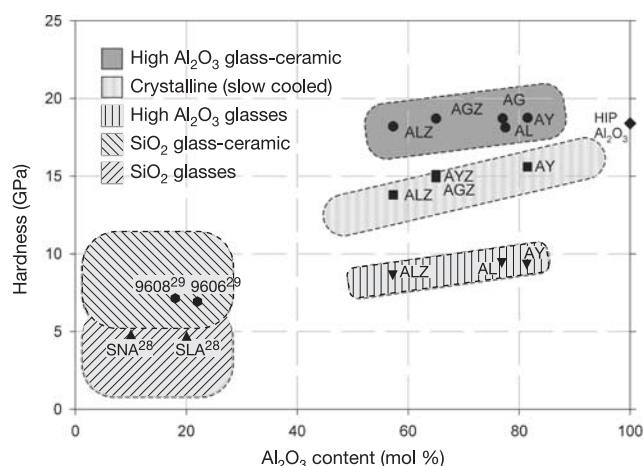


Figure 5 Map of hardness against Al_2O_3 content in different material classes. Properties of silica-based glasses (sodium aluminosilicate, lanthanum aluminosilicate) and glass-ceramics (Corning 9606 and 9608) were taken from refs 28 and 29. High-alumina glasses and glass-ceramics of this work are much harder than conventional silicates. They surpass other known oxides including BeO , MgO , Y_2O_3 , ZrO_2 , TiO_2 and $\text{Y}_3\text{Al}_5\text{O}_{12}$, and are comparable to pure $\alpha\text{-Al}_2\text{O}_3$ and $\beta\text{-Si}_3\text{N}_4$. Hot isostatically pressed Al_2O_3 was prepared in accordance with ref. 23. The compositions of the present study were also crystallized directly from the melt during slow cooling.

appreciable separation between T_g and T_x . Therefore, even partial densification requires very high pressures (more than 500 MPa)^{25,26} to activate dislocation-controlled, crystalline plasticity. In contrast, we have found that suitable kinetic windows do exist in the CaO–Al₂O₃, La₂O₃–TiO₂ and BaO–TiO₂ systems, which allowed us to perform bulk consolidation of rapidly quenched glass beads in a similar fashion to that for Al₂O₃–REO. On the basis of these results we believe that, as for bulk metallic glasses, the present glass-forming and consolidation approach is applicable to a large variety of ionic compositions and will open the door to many new bulk oxide glasses and nanocrystalline ceramics. □

Received 12 March; accepted 3 June 2004; doi:10.1038/nature02729.

1. Sun, K.-H. Aluminate glasses. *Glass Ind.* **30**, 199–200, 232 (1949).
2. Sarjeant, P. T. & Roy, R. in *Reactivity of Solids* (eds Mitchell, J. W., DeVries, R. C., Roberts, R. W. & Cannon, P.) 725–733 (Wiley, New York, 1969).
3. Weber, J. K. R., Abadie, J. G., Hixson, A. D., Nordine, P. & Jerman, G. A. Glass formation and polymorphism in rare-earth oxide-aluminum oxide compositions. *J. Am. Ceram. Soc.* **83**, 868–872 (2000).
4. Weber, R., Nordine, P. C., Key, T. & Tangeman, J. in *Proceedings of the SPIE – The International Society for Optical Engineering* Vol. 4990 (eds Shibin, J. & Jacques, L.) 70–76 (SPIE – Int. Soc. Opt. Eng., Bellingham, WA, 2003).
5. Liao, S.-C., Chen, Y.-J., Kear, B. H. & Mayo, W. E. High pressure/low temperature sintering of nanocrystalline alumina. *Nanostruct. Mater.* **10**, 1063–1079 (1998).
6. Chen, L.-W. & Wang, X.-H. Sintering dense nanocrystalline ceramics without final-stage grain growth. *Nature* **404**, 168–171 (2000).
7. Klement, W., Willens, R. & Duwez, P. Non-crystalline structure in solidified gold-silicon alloys. *Nature* **187**, 869–870 (1960).
8. Turnbull, D. & Cohen, M. H. Composition requirements for glass formation in metallic and ionic systems. *Nature* **189**, 131–132 (1961).
9. Rawson, H. *Inorganic Glass-forming Systems* 23–29 (Academic, New York, 1967).
10. Jantzen, C. M., Krepski, R. P. & Herman, H. Ultra-rapid quenching of laser-melted binary and unary oxides. *Mat. Res. Bull.* **15**, 1313–1326 (1980).
11. McMillan, P. & Piriou, B. Raman spectroscopy of calcium aluminate glasses and crystals. *J. Non-Cryst. Solids* **55**, 221–242 (1983).
12. Shelby, J. E. Formation and properties of calcium aluminosilicate glasses. *J. Am. Ceram. Soc.* **68**, 155–158 (1985).
13. Yajima, S., Okamura, K. & Shishido, T. Glass formation in the Ln–Al–O system (Ln: Lanthanoid and yttrium elements). *Chem. Lett. (Jpn.)* **8**, 1313–1326 (1980).
14. Kawamura, Y., Kato, H. & Inoue, A. Full strength compacts by extrusion of glassy metal powder at the supercooled liquid state. *Appl. Phys. Lett.* **67**, 2008–2013 (1995).
15. Waku, Y., Ohtsubo, H. & Inoue, A. A jelly-like ceramic fiber at 1193 K. *Mat. Res. Innovat.* **3**, 185–189 (2000).
16. Sarjeant, P. T. & Roy, R. Ti⁴⁺ coordination in glasses in RO–TiO₂ systems. *J. Am. Ceram. Soc.* **52**, 57–58 (1969).
17. Inoue, A., Zhang, T. & Masumoto, T. Glass-forming ability of alloys. *J. Non-Cryst. Solids* **156–158**, 473–480 (1993).
18. Shelby, J. E. & Kohli, J. T. Rare-earth aluminosilicate glasses. *J. Am. Ceram. Soc.* **73**, 39–42 (1990).
19. Angell, C. A. Structural instability and relaxation in liquid and glassy phases near the fragile liquid limit. *J. Non-Cryst. Solids* **102**, 205–221 (1988).
20. Beall, G. H. in *Advances in Nucleation and Crystallization in Glasses* (eds Hench, L. L. & Freiman, S. W.) 251–261 (American Ceramic Society, Columbus, Ohio, 1971).
21. Rosenflanz, A. *Fused Al₂O₃–Rare-Earth oxide–ZrO₂ Eutectic Materials*. US Patent 6,582,488 (2003).
22. Beall, G. H. & Pinckney, L. R. Nanophase glass-ceramics. *J. Am. Ceram. Soc.* **82**, 5–16 (1999).
23. Mizuta, H. et al. Preparation of high-strength and translucent alumina by hot isostatic pressing. *J. Am. Ceram. Soc.* **75**, 469–473 (1992).
24. Yip, S. Nanocrystals: the strongest size. *Nature* **391**, 532–533 (1998).
25. Kear, B. H., Liao, S.-C. & Mayo, W. E. *High Pressure and Low Temperature Sintering of Nanophase Ceramic Powders*. US Patent 6,395,214 (2002).
26. Freim, J., McKittrick, J., Nellis, W. J. & Katz, J. D. Development of novel microstructures in zirconia-toughened alumina using rapid solidification and shock compaction. *J. Mater. Res.* **11**, 110–119 (1996).
27. Wilding, M. C. & McMillan, P. F. Polymorphic transitions in yttria-alumina liquids. *J. Non-Cryst. Solids* **293–295**, 357–365 (2001).
28. Bansal, N. P. *Handbook of Glass Properties* Ch. 11 (Academic, New York, 1986).
29. McMillan, P. W. in *Glass-ceramics* (eds Roberts, J. P. & Popper, P.) 196–200 (Academic, New York, 1979).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Goodbrake (3M Company) for help with microscopy, and L.-W. Chen (University of Pennsylvania) for helpful comments during the preparation of the manuscript.

Competing interests statement The authors declare competing financial interests; details accompany the paper on www.nature.com/nature.

Correspondence and requests for materials should be addressed to A.R. (arosenflanz@mmm.com).

Self-assembly of amphiphilic dendritic dipeptides into helical pores

Virgil Percec¹, Andrés E. Dulcey¹, Venkatachalapathy S. K. Balagurusamy^{1,2}, Yoshiko Miura¹, Jan Smidrkal¹, Mihai Peterca^{1,2}, Sami Nummelin¹, Ulrica Edlund¹, Steven D. Hudson³, Paul A. Heiney², Hu Duan³, Sergei N. Magonov⁴ & Sergei A. Vinogradov⁵

¹Roy & Diana Vagelos Laboratories, Department of Chemistry, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6323, USA

²Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6396, USA

³National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8544, USA

⁴Digital Instruments, Veeco Metrology Group, Santa Barbara, California 93110, USA

⁵Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6059, USA

Natural pore-forming proteins act as viral helical coats¹ and transmembrane channels^{2–4}, exhibit antibacterial activity⁵ and are used in synthetic systems, such as for reversible encapsulation⁶ or stochastic sensing⁷. These diverse functions are intimately linked to protein structure^{1–4}. The close link between protein structure and protein function makes the design of synthetic mimics a formidable challenge, given that structure formation needs to be carefully controlled on all hierarchy levels, in solution and in the bulk. In fact, with few exceptions^{8,9}, synthetic pore structures capable of assembling into periodically ordered assemblies that are stable in solution and in the solid state^{10–13} have not yet been realized. In the case of dendrimers, covalent¹⁴ and non-covalent¹⁵ coating and assembly of a range of different structures^{15–17} has only yielded closed columns¹⁸. Here we describe a library of amphiphilic dendritic dipeptides that self-assemble in solution and in bulk through a complex recognition process into helical pores. We find that the molecular recognition and self-assembly process is sufficiently robust to tolerate a range of modifications to the amphiphile structure, while preliminary proton transport measurements establish that the pores are functional. We expect that this class of self-assembling dendrimers will allow the design of a variety of biologically inspired systems with functional properties arising from their porous structure.

The dendron (4-3,4-3,5)12G2-X with X = CO₂CH₃ and CO₂H, and the dendron (4-3,4-3,5)*n*G2-CH₂OH with *n* = 10–16, self-assemble into closed supramolecular columns¹⁸. The dendrons (4-3,4-3,5)12G2-CH₂-X containing as X-groups Boc(Moc)-L-Tyr-L-Ala-OMe, Boc-D-Tyr-D-Ala-OMe, Boc-L-Tyr-D-Ala-OMe, Boc-D-Tyr-L-Ala-OMe, Boc-DL-Tyr-DL-Ala-OMe and Moc-L-Tyr-L-OMe (Fig. 1a) were synthesized (Supplementary Schemes S1 and S2), and their dipeptide groups (short names for Boc-L-Tyr-L-Ala-OMe are L-Tyr-L-Ala or L-L dipeptide) were used as tags to monitor self-assembly and expression of chirality in solution and in bulk (Boc, *t*-butoxycarbonyl; Moc, methoxycarbonyl). In solvents allowing hydrogen-bonding, such as CHCl₃, CH₂Cl₂ and tetrahydrofuran (THF), self-assembly is not detected by temperature- and concentration-dependent ¹H-NMR, UV spectroscopy or circular dichroism (CD). This indicates that the equilibrium between the trans and gauche conformers of the benzyl ether moiety favours a globular dendron conformation encapsulating the dipeptide in its focal point (Fig. 1), thereby sterically prohibiting intermolecular hydrogen-bonding.

In contrast, hydrogen-bonding and self-assembly occur in the

solvophobic solvent cyclohexane: regardless of dipeptide stereochemistry, a decrease in temperature from 60 °C to 30 °C leads to a downfield shift of the NH protons of Tyr and Ala and of CH of Tyr (deshielding), and an upfield shift of the aromatic, benzyl and methylenic ether protons of the dendron, indicative of hydrogen-bonding and shielding effects due to intermolecular interactions (Fig. 2a). The chemical shifts of methyl, methoxy and CH groups of Ala are not affected. Below 30 °C, the ^1H -NMR spectrum broadens, corresponding to stiffening of the supramolecular structure. Except for the alkyl groups, no NMR spectrum is detected below 24 °C (Fig. 2a), indicating decreased molecular motion^{19,20}. On cooling from 60 °C to 42 °C, an increase in the absorption at 230 nm (A_{230}) of the UV spectrum (hyperchromic effect) and a blue shift are observed in cyclohexane (Fig. 2b). A_{230} is constant between 42 °C and 30 °C; upon cooling from 30 °C to 12 °C, a second blue shift of the A_{230} peak (hypochromic effect) and an isosbestic point are observed. The hypochromic effect¹⁹ is indicative of oriented chromophores, varying with their conformation and distance. Below 12 °C, the UV absorption spectrum shows no observable changes. Plotting A_{230} as a function of temperature yields a sigmoidal curve, indicative of a cooperative two-state process for self-assembly^{20,21}. In addition, the increase of A_{230} from 60 °C to 42 °C is associated with the transition from a globular dendron containing a mixture of trans and gauche benzyl ether conformers to an all trans-tapered dendron that facilitates self-assembly (Fig. 1). During this process, intramolecular interactions within the globular dendron are eliminated (causing the hyperchromic effect) and the trans-tapered dendron undergoes intermolecular hydrogen-bonding (Fig. 2a, b). Below 32 °C, the hypochromic and blue shift and the appearance of the isosbestic point indicate an equilibrium between the tapered dendron and its aggregate. At about 12 °C, this equilibrium is shifted entirely to the supramolecular aggregate.

CD experiments detect chirality in the supramolecular structure, and thus complement the ^1H -NMR and UV analysis. Chiral dipeptides in THF and chiral dendritic dipeptides in THF, CH_2Cl_2 , CHCl_3 and $\text{ClCH}_2\text{CH}_2\text{Cl}$ show in the CD spectra between 60 °C and 8 °C only the ellipticity of the dipeptide chromophore at $\lambda = 232$ nm, indicating molecular solutions (Supplementary Figs SF1 to SF3). Regardless of temperature and concentration, the CD of racemic (4-3,4-3,5)12G2-CH₂-(DL-Tyr-DL-Ala) does not exhibit any signal in cyclohexane or in other solvents (Supplementary Fig. SF2). From 60 °C to 30 °C, both the L-L and D-D dendritic dipeptides exhibit in cyclohexane only the ellipticity of the dipeptide (Fig. 2c, d and Supplementary Fig. SF3). These results, together with UV and NMR data (Fig. 2a, b), indicate that the dendron adopts an all-trans tapered conformation and is in fast exchange with its aggregate. Below 30 °C, only the trans-tapered conformer exists in molecular solution. In this temperature range, the transfer of chirality from the dipeptide to the aromatic part of the dendron, amplification of the Cotton effects and an isodichroic point are observed. This implies that the supramolecular aggregate must have an architecture equipped with a mechanism capable of amplifying the stereochemical information of the dipeptide. The plots of the chemical shifts (Fig. 2a) and ellipticity $[\theta]$ at 248 nm (Supplementary Fig. SF4) as a function of temperature reflect only the equilibrium between the trans-tapered dendron and its aggregate, while the UV analysis demonstrates the progression of the conformational order of the dendron. For the aggregate in cyclohexane (1.6×10^{-4} M), a transition temperature T_m of 22 °C is calculated from UV, CD and NMR plots as a function of temperature. For clarity, the NMR data in Fig. 2a are reported at a higher concentration (2.0×10^{-3} M) and show, as expected, a higher T_m (32 °C).

Differential scanning calorimetry (DSC) (Fig. 3a), small-angle powder X-ray diffraction (XRD) (Fig. 3b) and wide-angle oriented fibre XRD (Fig. 3c) show that the supramolecular structures assembled in bulk from (4-3,4-3,5)12G2-CH₂-(L-Tyr-L-Ala) and (4-3,4-3,5)12G2-CH₂-(D-Tyr-D-Ala) exhibit identical transition

temperatures and structure (Fig. 3a–c and Supplementary Table ST2). The CD spectra, tested for the absence of linear dichroism, and UV spectra of these dendrons recorded from cyclohexane solution or as thin films on quartz (Supplementary Fig. SF5) are identical except that L-L and D-D dendritic dipeptides show mirror-image Cotton effects (Fig. 2c, d). The similarity of the UV and CD spectra obtained from solution and in bulk indicates the same supramolecular structure in both states. The DL-DL, D-L and L-D dendritic dipeptides exhibit similar structures as the L-L and D-D (Fig. 3a–c) except that DL-DL, D-L and L-D dendritic dipeptides have slightly lower transition temperatures. Their helical sense is determined by the stereochemistry of Tyr. Thin-film CD spectra show this helical columnar structure in both ordered glassy and liquid crystalline states (Supplementary Fig. SF5). Although the DL-DL derived supramolecular structure is racemic, its fibre XRD demonstrates columns with short-range helical order (Fig. 3c). The presence of helicity in the columns of the racemic DL-DL dendritic

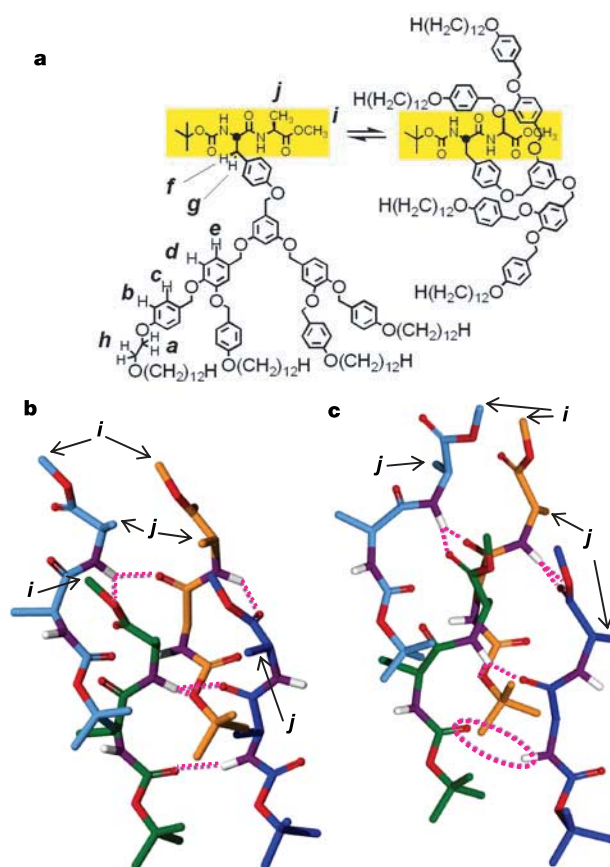


Figure 1 Structure, conformation and hydrogen-bonding of (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe and (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-D-Ala-OMe during self-assembly. **a**, Trans-tapered low-temperature (left), and globular high-temperature (right), conformers of the L-L stereoisomer. The letters (a, b, ..., j), some of them also indicated in **b** and **c**, indicate protons with ^1H -NMR chemical shifts marked in Fig. 2a. Yellow indicates the region of the dendritic dipeptide where functional groups capable of taking part in hydrogen-bonding are accessible (left), and the corresponding region that sterically prevents the same type of hydrogen-bonding in the globular structure (right). **b**, Hydrogen-bonding interactions of the L-L dipeptide forming the pore according to the model from Fig. 4e. **c**, Hydrogen-bonding interactions of the L-D dipeptide. In **b** and **c**, four dipeptides from two layers are illustrated (top layer has one peptide with carbons shown light blue and one with carbons shown gold, bottom layer has one peptide with carbons shown green and one with carbons shown dark blue; oxygen, red; hydrogen, white; nitrogen, purple; phenyl of Tyr and hydrogens not shown except in hydrogen-bonding N-H). The L-D dipeptide does not exhibit (dotted oval) the bottom hydrogen-bonding of the L-L dipeptide. This explains the lower transition temperature of the L-D dendritic dipeptide assembly (Fig. 3a). Hydrogen-bonding distances are in Supplementary Fig. SF14.

dipeptide suggests that helix conformation is induced by the achiral dendrons, with the chiral peptide attached to the dendron only selecting the twist sense of the helix. This assembly behaviour relates to other examples of stereocentres determining the twist sense of racemic helical structures^{20,21}, and contrasts with systems where a stereocentre induces helicity in an achiral non-helical supramolecular structure^{22,23}.

Transmission electron microscopy (TEM) images and electron diffraction (ED) patterns (Fig. 3e) of the homeotropically aligned, and scanning force microscopy (SFM) images of the planar-aligned, two-dimensional hexagonal columnar lattice (Fig. 3f) confirm and complement the XRD analysis. Moreover, TEM images and their Fourier reconstructions show columns with low electron density both in the core and in their aliphatic periphery (Fig. 3e). These images contrast with previously observed TEM for closed-core columns with a high electron density in their core¹⁷. The low electron density in the core is associated with the hollow structure of the cylinder, and explains the anomalously enhanced intensities of the higher-order diffraction peaks (11), (20) and (21) observed in XRD (Fig. 3b) and ED (Fig. 3e). Absolute electron density profiles were computed from the XRD data (Supplementary Figs SF10 and SF11) assuming an intramolecular phase segregated column^{14,24}. The phases of the (10), (11) and (20) reflections were established from the Fourier analysis of the TEM images (Supplementary Figs SF7 and SF8), and those of higher order reflections (21) and (30) were phase combinations that resulted in nearly constant electron density for the aromatic and aliphatic regions satisfying the intramolecular segregated model^{17,24}. The converted electron density profiles (Supplementary Figs SF10 and SF11) of the assembly exhibit significantly lower electron density in the core than the average aliphatic density (approximately 0.30 electrons Å⁻³ for a mass density of 0.86 g cm⁻³) in the periphery and demonstrate hollow columns (Fig. 3d). Form-factor calculations were performed for a column model with three levels of electron density (hollow core, high density peptide-aromatic region, low density aliphatic

periphery) distribution. These calculations, in combination with experimental densities, electron density profiles and molecular modelling experiments, were performed for the entire series of supramolecular structures generated from L-L dendritic dipeptides with $n = 6, 8, 10, 12, 14$ and 16 to determine their pore diameters (D_{pore}) and structure. The external diameters (D_{ext}) were calculated from the peak positions in the XRD^{15,18} (Fig. 3b). Values of D_{pore} were obtained from the least-squares fit of the diffraction amplitudes calculated from the three level electron density model to the measured XRD amplitudes. The dendritic dipeptides with $n = 12$ and D-D, D-L, L-L and DL-DL stereochemistry have D_{pore} values of 13.6, 12.8, 13.7 and 12.8 Å.

The side view of the supramolecule generated from (4-3,4-3,5)12G2-CH₂-(L-Tyr-L-Ala), its top view (only methyl groups are shown as the alkyl groups of the dendron), the top view of a single layer of the pore, and the cross-section of the pore without the dendron, are shown in Fig. 4a–d. A right-handed column is self-assembled from the L-L dendritic dipeptide and a left-handed one from the D-D stereoisomer. The pore interior is hydrophobic, containing the methyl group of Ala (white) and one methyl group of the Boc (blue) on the pore surface. The hydrophilic part of the dipeptide is segregated between the hydrophobic dendron and the hydrophobic pore. Hydrophobic channels are important, because they facilitate the transport of ions² and water^{3,25} with both high rate and selectivity. The conformation of the dendritic dipeptide and the hydrogen-bonding interactions that generate the supramolecular assembly and the inner part of the pore are illustrated in Fig. 1b, c and Supplementary Figs SF13 and SF14. The dipeptide forms an interdigitated and hydrogen-bonded β -helix²⁶ in response to the self-assembly of its dendritic fragment. The structure of the pore resembles a β -barrel^{7,13}. The classic antiparallel dipeptide model¹³ was eliminated first, as it does not form a column. Eight additional structures (four non-helical and four helical) of pore assembly were considered (Supplementary Fig. SF12) before selecting the model shown in Fig. 4e, which is stabilized via a hydrogen-bonding

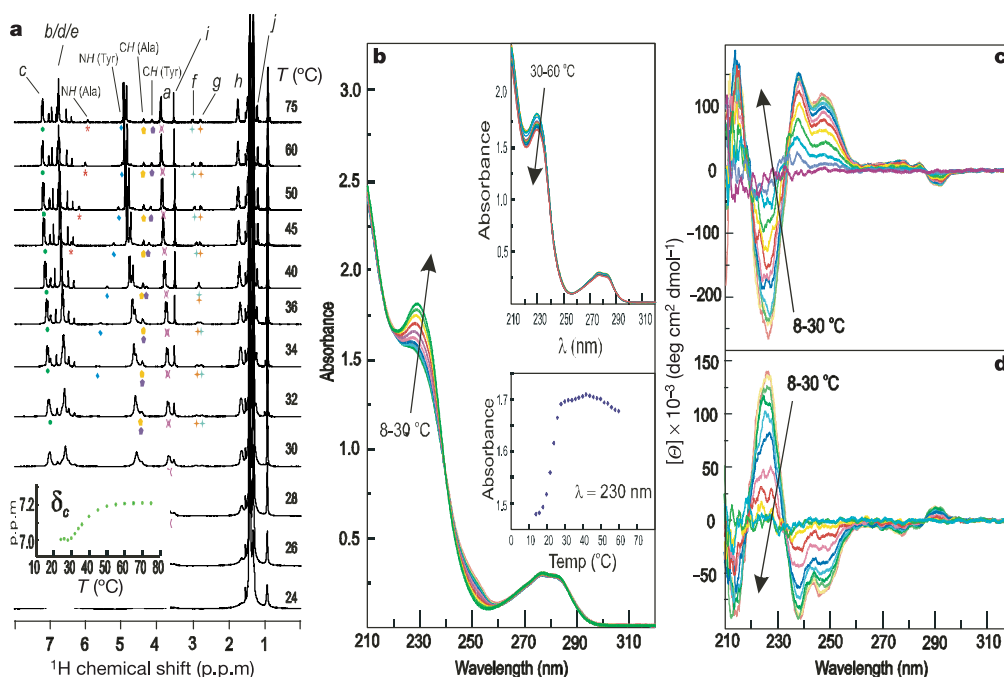


Figure 2 Spectroscopic analysis of dendritic dipeptide self-assembly in solvophobic solution. **a**, ¹H-NMR (500 MHz) spectra of (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe in C₆D₁₂ (2.0×10^{-3} M). Inset shows the plot of chemical shift, δ_c , as a function of temperature. Proton assignments are shown in Fig. 1. **b**, UV spectra of (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe in cyclohexane (1.6×10^{-4} M) exhibiting an isosbestic point

at 240 nm. Insets are spectra at higher temperature and a plot of A_{230} as a function of temperature. **c**, CD spectra of (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe in cyclohexane (1.6×10^{-4} M). **d**, CD spectra of (4-3,4-3,5)12G2-CH₂-Boc-D-Tyr-D-Ala-OMe in cyclohexane (1.6×10^{-4} M). In all parts, arrows indicate trends upon increasing temperature.

network (Fig. 1b, c). This model is unrelated to that of the single crystal structure of the dipeptide without dendron (Supplementary Fig. SF15).

Figures 1b, c, Fig. 4 and Supplementary Fig. SF14 suggest methods to redesign the pore architecture by using the retro-structural analysis of its XRD generated structure^{15,18}. For example, the replacement of Boc from (4-3,4-3,5)12G2-CH₂-(L-Tyr-L-Ala) with Moc reduced the pore diameter from 12.8 Å to 10.2 Å. The conformation of the phenyl from Tyr is anti with respect to its Boc group, and is tilted in the opposite direction (Supplementary Fig. SF13a). Therefore, attaching the dendron to Tyr via an ester rather than benzyl ether was expected to restrict its dynamics; this has been demonstrated by the structure of (4-3,4-3,5)12G2-CO₂-(L-

Tyr-L-Ala), which, despite $D_{\text{pore}} = 12.4 \text{ Å}$, requires longer annealing to achieve structural order as measured by XRD. The replacement of L-Ala from (4-3,4-3,5-)12G2-CH₂-(L-Tyr-L-Ala) with other non-polar or polar amino acids such as Gly, L-Val, L-Phe and L-Ser, of L-Tyr with L-Cys, of (L-Tyr-L-Ala) with (L-Ala-L-Tyr), and of (4-3,4-3,5)12G2- with other dendrons produced building blocks that self-assemble into related porous columns (Supplementary Tables ST6 and ST8). The smaller D_{pore} of (4-3,4-3,5)12G2-CH₂-(Boc-L-Cys-L-Ala-OMe) (10.0 Å versus 13 Å for the case of the same dendritic dipeptide with Boc-L-Tyr-L-Ala-OMe dipeptide) suggests that the phenyl group from Tyr plays an important role in the self-assembly of porous columns. Therefore, a library of self-assembling dendrons containing additional phenyl and biphenyl groups in their apex and/or branches was designed to produce 20 non-peptidic porous columns with D_{pore} ranging from 2 to 24 Å (Supplementary Tables ST9 to ST13). These structures, together with those generated from the 19 dendritic dipeptides (Supplementary Tables ST5 to ST8), demonstrate the simplicity and versatility of this approach to the fabrication of both non-biological pores and tubular liquid crystals^{10,11}.

Successful self-assembly of the dendritic dipeptides in cyclohexane suggests that they may also assemble in phospholipid membranes, and suitable decoration of their alkyl groups might mediate their assembly in or on the surface of microbial cell membranes⁵. As first step in this direction, the porous structure of (4-3,4-3,5)12G2-

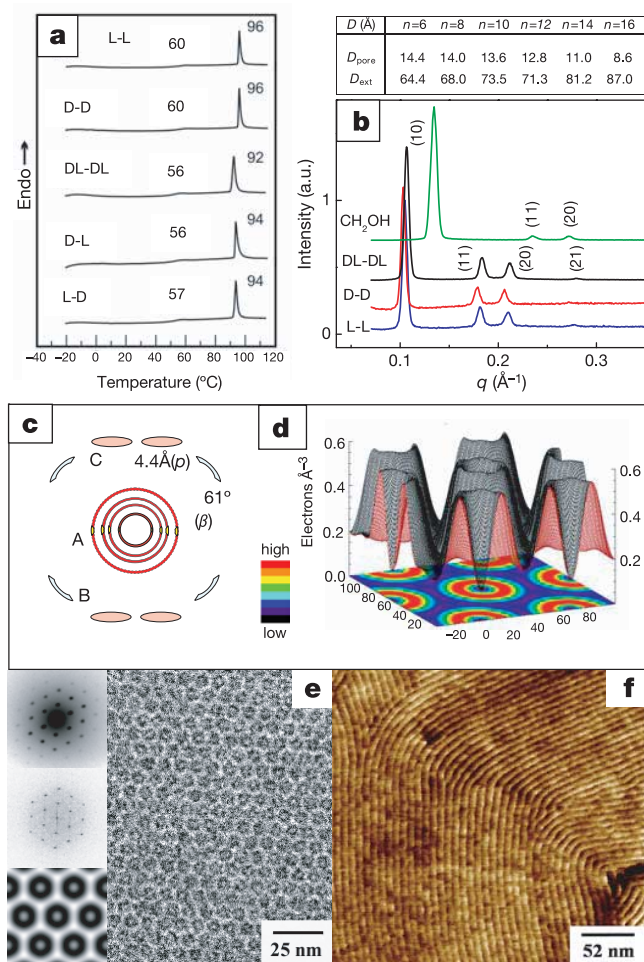


Figure 3 Structural analysis of dendritic dipeptide pore in bulk. **a**, DSC showing the glassy and isotropization temperatures of L-L, D-D, DL-DL, D-L and L-D stereoisomers of (4-3,4-3,5)12G2-CH₂-Boc-Tyr-Ala-OMe. **b**, Powder XRD of L-L, D-D and DL-DL stereoisomers of (4-3,4-3,5)12G2-CH₂-Boc-Tyr-Ala-OMe and of (4-3,4-3,5)12G2-CH₂OH. Also shown (top table) are D_{ext} and D_{pore} (Å) of (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe with $n = 6$ to 16. **c**, Wide and small-angle fibre XRD pattern. A, column to column distance, long-range order; B, molecular tilt (β); C, short-range helical correlation along column axis, short-range pitch (p). **d**, Electron density maps of (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe columns. Profile shows variation of electron density in a plane perpendicular to columns. Coloured contour maps show change in electron density in the same plane (x and y axes of the plane are in Angstroms). **e**, TEM of (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe along the column axis. Insets from top are: electron diffraction pattern, Fourier transform power spectrum, and image reconstructed from the (10), (11) and (20) Fourier components with phases: +, -, -. **f**, SFM of columns parallel to pyrolytic graphite substrate. Dislocations and disclinations of the focal conic like texture are observed.

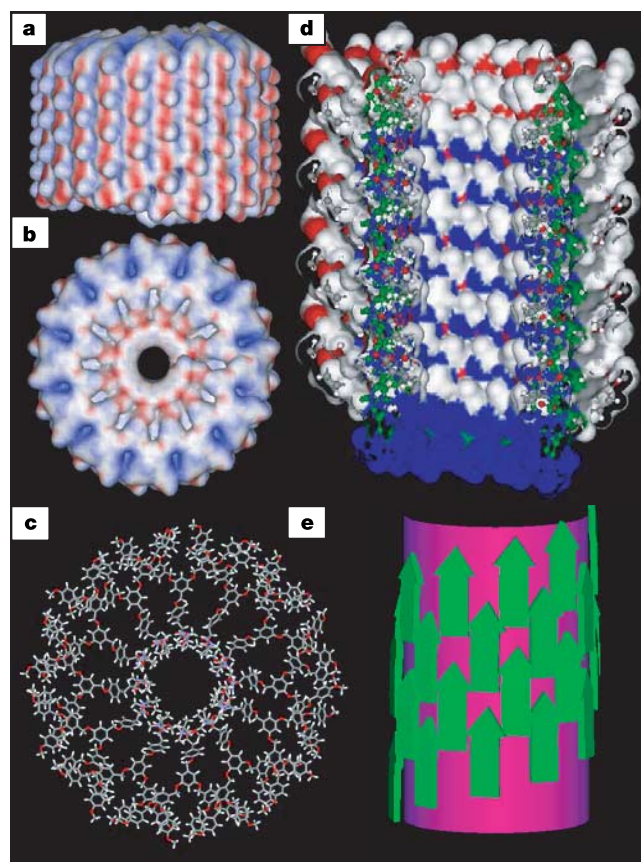


Figure 4 Molecular models of the helical porous columns self-assembled from (4-3,4-3,5)12G2-CH₂-Boc-L-Tyr-L-Ala-OMe (for simplicity $n = 12$ was replaced with $n = 1$). **a**, Side view of the right-handed column. **b**, Top view of a single porous column layer. **c**, Top view of a single porous column layer. **d**, Cross-section through the hydrophobic pore (without dendrons) showing its β -barrel structure (CH₃ of Ala is white, CH₃ of Boc are blue, O is red, C-N of dipeptide are green, aromatic groups are grey) assembled from the β -helical dipeptides. **e**, Schematic model for the self-assembly of the dipeptidic β -barrel pore. The green arrows indicate the dipeptides.

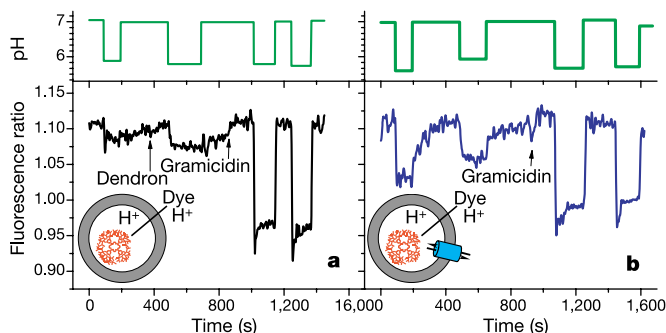


Figure 5 Proton transport through (4-3,4-3,5)12G2-CH₂-(Boc-L-Tyr-L-Ala-OMe) pores reconstituted in phospholipid liposomes (pH-jump experiments). **a**, Liposomes containing only the membrane-impermeable pH indicator²⁸ inside. **b**, Liposomes containing the pH indicator inside and the dendritic dipeptide pores. In both cases, arrows indicate the addition of the dendritic dipeptide or gramicidin as DMSO/THF solutions. pH jumps at 20 °C outside the liposome (induced by adding aliquots, about 10 μ l, of HCl or KOH) were recorded by pH microelectrodes (upper graphs). pH jumps inside liposomes were assessed by fluorescence (I_{647}/I_{670}) (lower graphs). The signal of the total amount of captured pH dye was estimated by adding an excess of gramicidine. Liposomes were prepared by sonicating a 1/14 mass ratio of dendritic dipeptide in the presence of L- α -phosphatidylcholine (P5638 from Sigma) and a fluorescent membrane-impermeable pH indicator (G4 polyglutamic porphyrin-dendrimer)²⁸ in a phosphate buffer (10 mM K₂HPO₄, 50 mM KCl, pH = 7.0). The control experiment (**a**) has no dendron. Liposomes were purified from untrapped indicator by gel filtration on Sephadex (G200) and on anion exchange resin QAE Sepharose A50 and placed in a fluorimetric cell equipped with a stirrer. As expected from its hydrophobicity (un-optimized experiment **a**), the dendritic dipeptide was not delivered very effectively to liposomes by simply adding its solution in DMSO/THF: the addition increases permeability only slightly. In contrast, liposomes made of a lipid dendritic dipeptide mixture²⁷ (14/1 mass ratio lipid to dendron equivalent to an average of one to two pores per vesicle; Supplementary Section S10) yields permeable vesicles significantly more responsive to pH changes (**b**). Addition of gramicidin increases the magnitude of the jumps, suggesting that a small fraction of vesicles did not contain dendritic channels. Addition of 10 μ l of DMSO alone does not affect permeability.

CH₂-(Boc-L-Tyr-L-Ala-OMe) was reconstituted²⁷ in the thermotropic bilayer lamellar phase and in liposomes produced from phospholipids (Supplementary Section S9). By monitoring the emission intensity of a pH-sensitive fluorescent dye²⁸ captured inside the liposomes, proton translocation mediated by dendritic pores and gramicidin channels can be evaluated (Fig. 5 and Supplementary Fig. SF19)^{28,29}. Proton permeability of liposomes containing an average of one to two reconstituted dendritic pores (14/1 mass ratio phospholipid to dendritic dipeptide) was comparable in efficiency to those containing gramicidin channels. These results illustrate that supramolecular dendrimer chemistry³⁰ allows the controlled design of a range of periodic non-biological porous structures forming in solution and as films. $\square$

Received 11 February; accepted 16 June 2004; doi:10.1038/nature02770.

- Klug, A. From macromolecules to biological assemblies. *Angew. Chem. Int. Edn Engl.* **22**, 565–582 (1983).
- Doyle, D. A. *et al.* The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity. *Science* **280**, 69–77 (1998).
- Murata, K. *et al.* Structural determinants of water permeation through aquaporin-1. *Nature* **407**, 599–605 (2000).
- van den Berg, B. *et al.* X-ray structure of a protein conducting channel. *Nature* **427**, 36–44 (2004).
- Fernandez-Lopez, S. *et al.* Antibacterial agents based on the cyclic D,L- α -peptide architecture. *Nature* **412**, 452–455 (2001).
- Ishii, D. *et al.* Chaperonin-mediated stabilization and ATP-triggered release of semiconductor nanoparticles. *Nature* **423**, 628–632 (2003).
- Bayley, H. & Cremer, P. S. Stochastic sensors inspired by biology. *Nature* **413**, 226–230 (2001).
- Ghadiri, M. R., Granja, J. R., Milligan, R. A., McRee, D. E. & Khazanovich, N. Self-assembling organic nanotubes based on a cyclic peptide architecture. *Nature* **366**, 324–327 (1993).
- Schmitt, J.-L., Stadler, A.-M., Kyritsakas, N. & Lehn, J.-M. Helicity encoded molecular strands: efficient access by the hydrazone route and structural features. *Helv. Chim. Acta* **86**, 1598–1624 (2003).
- Lehn, J.-M. *Supramolecular Chemistry: Concepts and Perspectives* 118 (VCH, Weinheim, 1995).

- Bong, D. T., Clark, T. D., Granja, J. R. & Ghadiri, M. R. Self-assembling organic nanotubes. *Angew. Chem. Int. Edn Engl.* **40**, 989–1011 (2001).
- Hill, D. J., Mio, M. J., Prince, R. B., Hughes, T. S. & Moore, J. S. A field guide to foldamers. *Chem. Rev.* **101**, 3893–4011 (2001).
- Sakai, N. & Matile, S. Synthetic multifunctional pores: lessons from rigid-rod β -barrels. *Chem. Commun.*, 2514–2523 (2003).
- Percec, V. *et al.* Controlling polymer shape through the self-assembly of dendritic side-groups. *Nature* **391**, 161–164 (1998).
- Percec, V. *et al.* Self-organization of supramolecular helical dendrimers into complex electronic materials. *Nature* **419**, 384–387 (2002).
- Percec, V., Cho, W.-D., Ungar, G. & Yeardley, D. J. P. Synthesis and NaOTf mediated self-assembly of monodendritic crown-ethers. *Chem. Eur. J.* **8**, 2011–2025 (2002).
- Hudson, S. D. *et al.* Direct visualization of individual cylindrical and spherical dendrimers. *Science* **278**, 449–452 (1997).
- Percec, V., Cho, W.-D., Ungar, G. & Yeardley, D. J. P. Synthesis and structural analysis of two constitutional isomeric libraries of AB₂-based monodendrons and supramolecular dendrimers. *J. Am. Chem. Soc.* **123**, 1302–1315 (2001).
- Nelson, J. C., Saven, J. G., Moore, J. S. & Wolynes, P. G. Solvophobically driven folding of nonbiological oligomers. *Science* **277**, 1793–1796 (1997).
- Brunsveld, L., Zhang, H., Glasbeek, M., Vekemans, J. A. J. M. & Meijer, E. W. Hierarchical growth of chiral self-assembled structures in protic media. *J. Am. Chem. Soc.* **122**, 6175–6182 (2000).
- Chan, H. S., Bromberg, S. & Dill, K. A. Models of cooperativity in protein folding. *Phil. Trans. R. Soc. Lond. B* **348**, 61–70 (1995).
- Engelkamp, H., Middelbeek, S. & Nolte, R. J. M. Self-assembly of disk-shaped molecules to coiled-coil aggregates with tunable helicity. *Science* **284**, 785–788 (2001).
- Hirschberg, J. H. K. K. *et al.* Helical self-assembled polymers from cooperative stacking of hydrogen-bonded pairs. *Nature* **407**, 167–170 (2000).
- Balagurusamy, V. S. K., Ungar, G., Percec, V. & Johansson, G. Rational design of the first spherical supramolecular dendrimers self-organized in a novel thermotropic cubic liquid-crystalline phase and the determination of their shape by X-ray analysis. *J. Am. Chem. Soc.* **119**, 1539–1555 (1997).
- Hummer, G., Rasaiah, J. C. & Noworita, J. P. Water conduction through the hydrophobic channel of a carbon nanotube. *Nature* **414**, 188–190 (2001).
- Cornelissen, J. J. L. M. *et al.* β -Helical polymers from isocyanopeptides. *Science* **293**, 676–680 (2001).
- Rigaud, J.-L., Pitard, B. & Levy, D. Reconstitution of membrane proteins into liposomes: application to energy-transducing membrane proteins. *Biochim. Biophys. Acta* **1231**, 223–246 (1995).
- Finikova, O. *et al.* Porphyrin and tetrabenzoporphyrin dendrimers: tunable membrane-impermeable fluorescent pH nanosensors. *J. Am. Chem. Soc.* **125**, 4882–4893 (2003).
- Ghadiri, M. R., Granja, J. R. & Buehler, L. K. Artificial transmembrane ion channels from self-assembling peptide nanotubes. *Nature* **369**, 301–304 (1994).
- Emrick, T. & Fréchet, J. M. J. Self-assembly of dendritic structures. *Curr. Opin. Colloid Interface Sci.* **4**, 15–23 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements Financial support by the National Science Foundation, the Office of Naval Research and the P. Roy Vagelos Chair at the University of Pennsylvania is acknowledged. J.S. thanks the Isabel and Alfred Bader Foundation for a graduate fellowship, and U.E. acknowledges a Hans Werthén scholarship for postdoctoral studies. We also thank S.Z.D. Cheng for density measurements, and G. Ungar for reading the draft manuscript and for suggestions.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.P. (percec@sas.upenn.edu).

Quantification of modelling uncertainties in a large ensemble of climate change simulations

James M. Murphy¹, David M. H. Sexton¹, David N. Barnett¹, Gareth S. Jones¹, Mark J. Webb¹, Matthew Collins¹ & David A. Stainforth²

¹Hadley Centre for Climate Prediction and Research, Met Office, FitzRoy Road, Exeter EX1 3PB, UK

²Department of Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK

Comprehensive global climate models¹ are the only tools that account for the complex set of processes which will determine future climate change at both a global and regional level. Planners are typically faced with a wide range of predicted changes from different models of unknown relative quality^{2,3}, owing to large but unquantified uncertainties in the modelling

process⁴. Here we report a systematic attempt to determine the range of climate changes consistent with these uncertainties, based on a 53-member ensemble of model versions constructed by varying model parameters. We estimate a probability density function for the sensitivity of climate to a doubling of atmospheric carbon dioxide levels, and obtain a 5–95 per cent probability range of 2.4–5.4 °C. Our probability density function is constrained by objective estimates of the relative reliability of different model versions, the choice of model parameters that are varied and their uncertainty ranges, specified on the basis of expert advice. Our ensemble produces a range of regional changes much wider than indicated by traditional methods based on scaling the response patterns of an individual simulation^{5,6}.

Detailed and multi-variable predictions of anthropogenic climate change are required for impact assessments⁶. Only comprehensive three-dimensional global climate models¹ (GCMs) are capable of providing such information. By simulating the key physical processes involved, they can represent the complex nonlinear interactions that influence climate change at a regional level, including changes in the frequency of damaging storms and other extreme events⁷. Yet GCM predictions are currently subject to considerable uncertainties in the modelling process^{4,8}, to the extent that different models often disagree even on the sign of the changes expected in particular regions². It is therefore essential that GCM predictions are accompanied by quantitative estimates of the associated uncertainty in order to render them usable in planning mitigation and adaptation strategies⁶.

Modelling uncertainties arise from fundamental choices made when building the GCM (for example, grid resolution), and from the parameterization of processes unresolved at the grid scale (for example, cloud formation). Here we provide a systematic investigation of modelling uncertainty by varying the settings of GCM parameters whose values cannot be accurately determined from observations⁹. The range of predictions obtained should be recognized as a lower limit, which could increase once the approach is generalized to sample structural modelling uncertainties¹⁰ arising from choices such as resolution, the set of processes included in the model and the basic assumptions on which its parameterizations are based¹¹.

We focus on uncertainties in the equilibrium response to a doubling of atmospheric CO₂ assuming no changes in other external forcing agents. Our GCM consists of the atmospheric model HadAM3¹² coupled to a mixed layer ocean that allows integration to equilibrium in a few decades. This enables us to increase ensemble size at the expense of neglecting ocean circulation feedbacks. There are of the order of 100 parameters in HadAM3, consisting of logical switches or variable coefficients or thresholds. A subset of 29 parameters was identified by modelling experts as controlling key physical characteristics of sub-grid scale atmospheric and surface processes. We perturbed these one at a time relative to the standard version of the GCM¹² (hereafter STD), creating a perturbed physics ensemble (PPE) of 53 model versions each used to simulate present-day and doubled CO₂ climates.

Uncertainty in regional climate change is commonly estimated by scaling a pattern of change from a single GCM simulation according to a range of possible changes in globally averaged surface temperature⁶. We assess this approach by synthesizing an ensemble of response patterns, in which the pattern of STD is scaled using 52 ratios of the climate sensitivity of each PPE member relative to STD. The ensemble standard deviations of synthesized and simulated responses are then compared (the black and red curves in Fig. 1). Scaling the pattern of a single ensemble member fails to reproduce the range of simulated regional changes, particularly for variables other than surface temperature (we show changes in precipitation and sea level pressure). For example, the scaling approach captures less than 10% of the variance of tropical precipitation changes. This

has fundamental implications for the way in which climate predictions are produced. A single prediction of future climate made with even the most sophisticated GCM will be of limited use for impact assessments. Only large ensembles of GCM predictions sampling the widest possible range of modelling uncertainties can provide a reliable specification of the spread of possible regional changes.

The spread of changes obtained from the PPE is influenced by both process uncertainties and internal variability arising from random climate variations¹³. We find that the spread in surface warming predictions is dominated by process uncertainties (compare red and blue curves in Fig. 1). For precipitation and sea level pressure, internal variability makes a much larger contribution, particularly in extratropical regions. Narrowing the uncertainties arising from internal variability will require ensembles of simulations started from different initial conditions, while the minimization of process uncertainties will require the development of more accurate parameterizations for use in GCMs.

The PPE also allows us to obtain credible quantitative estimates of the robustness of the simulated changes to the modelling uncertainties explored (Fig. 2). Robustness is high almost everywhere for surface temperature, but varies widely with location for precipitation and sea level pressure. The type of information shown in Figs 1 and 2 provides a basis for constructing climate scenarios in which the signal of expected regional change can be reliably

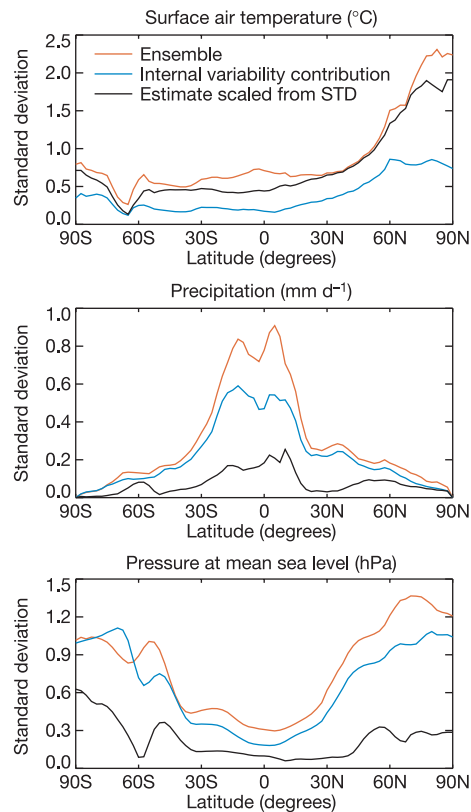


Figure 1 Zonal means of the ensemble standard deviation at individual 300 × 300 km² model grid boxes of the equilibrium response to doubled CO₂. Values are shown for 20 yr averages of December to February climate for surface air temperature, precipitation and pressure at mean sea level. The red curve shows results from the perturbed physics ensemble (PPE), the spread in which arises from both process uncertainties and internal variability. The blue curve shows the spread arising from internal variability alone, estimated by running 600 yr present-day and doubled CO₂ integrations of STD (the standard model version) and calculating the standard deviation of the response from constituent 20 yr periods. The black curve shows an attempt to reproduce the red curve by scaling the response patterns of STD according to the global climate sensitivities of the other 52 members of the PPE as described in the text.

separated from the noise of natural variability, thus providing planners with an improved basis for the development of appropriate response strategies⁶. We emphasize that the spread of regional changes estimated from our PPE, while substantially larger than would have been inferred by scaling the patterns of one of its members (see above), is likely to increase once it becomes possible to sample modelling uncertainties more comprehensively.

An important benchmark of anthropogenic climate change is the climate sensitivity, defined as the equilibrium response of globally averaged annual surface temperature to doubled CO₂. We estimate probability density functions (PDFs) for climate sensitivity from the 53-member PPE by assuming that the impacts of individual parameter perturbations, both on simulated present-day climate and the feedbacks that determine sensitivity, combine linearly. This allows us to predict the results of a much larger ensemble containing 4×10^6 model versions with randomly chosen multiple parameter perturbations generated by assuming a uniform distribution for each parameter within the range of values estimated by experts (see Methods). The blue curve in Fig. 3 shows a PDF obtained by assuming all 4×10^6 predictions of sensitivity to be equally reliable. It gives a median value of 2.9 °C with a spread (corresponding to a 5–95% probability range) of 1.9–5.3 °C.

The assumption of equal reliability between members of the

GCM ensemble is standard in climate prediction^{4,14}, yet represents a major difficulty since variations in quality between models are ignored. Here we introduce a Climate Prediction Index (CPI), an objective measure of reliability that can be used to weight different GCMs according to the estimated relative likelihood that they will correctly predict climate change in the real world¹⁵. Reliability can potentially be quantified by verifying simulations of climate change during the past century^{10,16}, or of a stationary climate assumed to correspond to some recent period^{17,18}. Our experimental design precludes the former approach, so we base the CPI on a broad range of present-day climate variables (Fig. 4 and Supplementary Information). Most PPE members occupy a rather narrow range of overall CPI values, though the range is much wider for some of its components, notably those associated with cloud, radiation and moisture. We justify use of the CPI to weight climate change predictions from a ‘perfect model’ test in which pairs of PPE members are compared against each other, one member taken to represent the observed climate system and the other a model simulation of it. For each possible pair, we calculated a CPI score and the magnitude of the difference between the simulated and ‘observed’ climate sensitivity. Amongst poor predictions of the ‘observed’ response (differences in sensitivity above the median) the CPI score was 2.7 times more likely to be poor (that is, above its median value) than good, and vice versa for good predictions of the response (sensitivity differences below the median).

We produce a likelihood-weighted PDF of climate sensitivity (the red curve in Fig. 3) by estimating the CPI of the 4×10^6 model versions used to produce the blue curve, and weighting their predictions of sensitivity according to $\exp(-0.5\text{CPI}^2)$ (see Methods). This results in a narrowing of the 5–95% probability range to 2.4–5.4 °C, while the median value increases to 3.5 °C. Previously, PDFs of climate sensitivity have been obtained by exploring the range of predictions of simpler climate models^{19–21} consistent with uncertainties in observed transient climate change and forcing^{22,23}. Our PDFs are (to our knowledge) the first to be determined by systematically exploring uncertainties in the complex variety of processes that actually determine climate sensitivity. They indicate a smaller probability for sensitivities of 2 °C or less than is implied by studies comparing observed historical changes with simulations by simple models²¹ or GCMs²⁴. Our PDFs are contingent upon the structural choices made in building our GCM, the use of a linear prediction scheme, the choice and application of observational constraints and the choice of parameters for perturbation. They also depend on the assumed distributions of parameter values, although we found that increasing their expert-specified ranges had only a modest impact on the 5–95% probability range

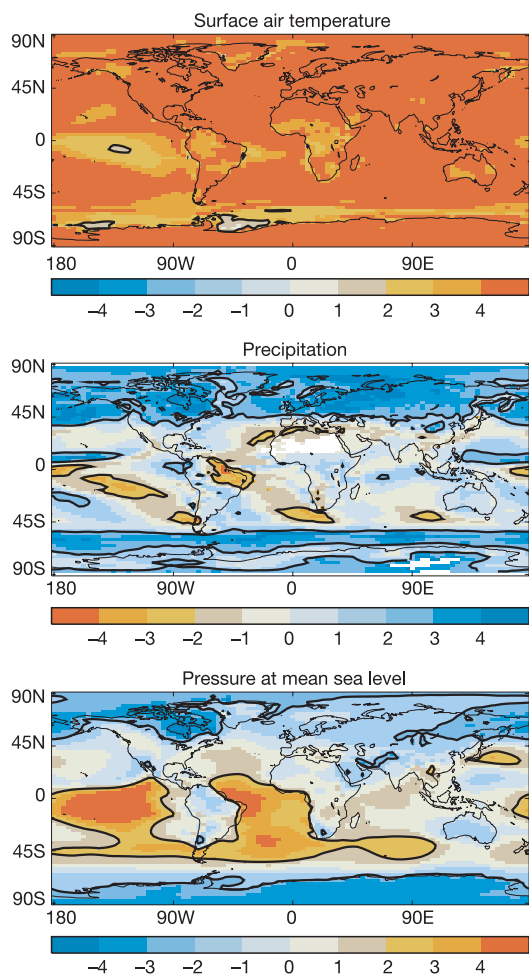


Figure 2 Robustness of simulated changes in surface air temperature, precipitation and pressure at mean sea level in response to doubled CO₂. The maps show changes in 20 yr means of December to February climate, averaged over the PPE of GCM versions and divided by the ensemble standard deviation of the changes. Values outside the range ± 2 (highlighted by the black contour) are taken to indicate a robust response.

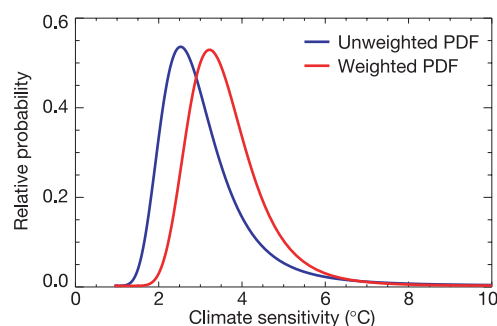


Figure 3 Probability distributions of climate sensitivity. These were obtained using linear statistical estimation of GCM predictions likely to result from a large PPE designed to sample the model parameter space comprehensively, with (red) and without (blue) weighting according to the estimated reliability of model versions based on the Climate Prediction Index (CPI). Details in text.

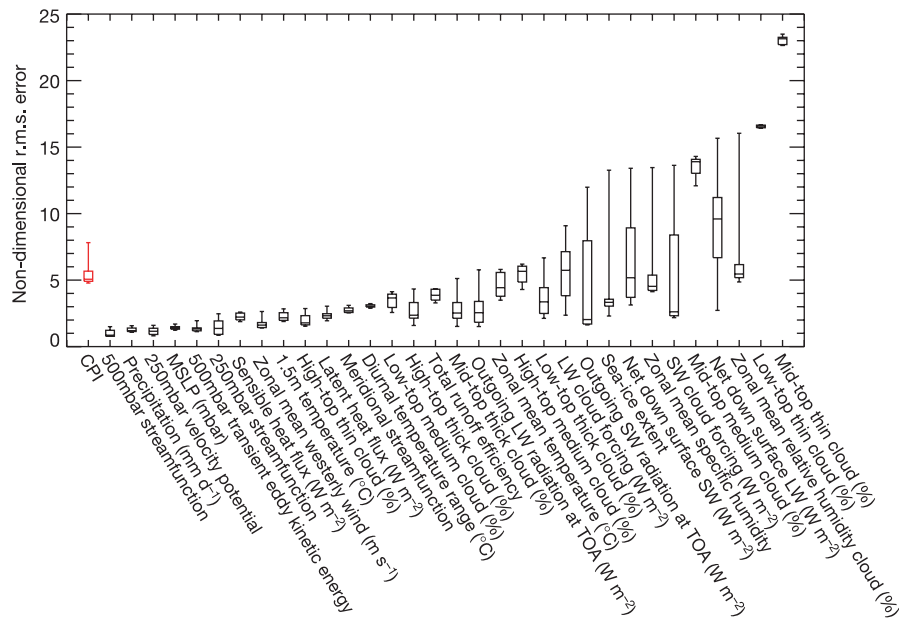


Figure 4 Values of the Climate Prediction Index (CPI) (red box and bars) and its components (black boxes and bars) from the PPE. The 32 components represent surface and atmospheric variables, and are calculated as the r.m.s. difference between simulated and observed present-day climatological mean patterns divided by the r.m.s. value of the standard deviation of simulated interannual variations. The plot shows averages of values calculated separately for each season of the year. Bars show the full range of the ensemble distribution of values, boxes show the range encompassed by the 5th and 95th

percentiles, and the horizontal line within each box shows the median. The CPI is calculated as the r.m.s. value of the 32 components for a given ensemble member. All components are weighted equally, apart from the nine fields of cloud cover which receive a relative weight of 1/3 since observations of high, medium and low cloud are interdependent for a given thickness category. LW, longwave; SW, shortwave; TOA, top of atmosphere; MSLP, mean sea level pressure.

associated with our CPI-weighted PDF (Supplementary Information). Our experimental design does not sample ocean circulation feedbacks or the impact of biases in the present-day simulations of sea surface temperature (SST). However these are likely to exert only a modest influence on global climate sensitivity (refs 25–28 and Supplementary Information). The impact of neglecting structural modelling uncertainties cannot yet be quantified, however the range obtained from our unweighted PDF encompasses the range of climate sensitivities (2.0–5.1°C) found in an ensemble of 15 GCMs developed at different modelling centres and containing structural variations⁴.

The PPE approach will be further developed to produce PDFs of time-dependent regional changes for use in assessments of climate-related risks⁶. This will require ensemble simulations of twentieth and twenty-first century climate using versions of HadAM3 coupled to a comprehensive ocean component¹, allowing us to account for the effects of oceanic thermal inertia, circulation changes and process uncertainties. These ensembles will need to sample multiple parameter perturbations, since our assumption that individual perturbations combine linearly is unlikely to be valid at a regional scale²⁹. Ensemble size will be increased by including results from simulations run on personal computers owned by members of the public and businesses (see refs 29, 30, and <http://www.climateprediction.net>). We also encourage other climate modelling institutes to perform similar ensemble experiments with their GCMs. These could then be combined with ours to create ‘super-ensembles’ that sample structural uncertainties. □

Methods

GCM integrations and parameter perturbations

For each ensemble member, control (that is, present day) and doubled CO₂ GCM integrations were run to equilibrium followed by a further 20 yr from which climate statistics were generated. The GCM used a mixed layer ocean with prescribed heat transports, which ensured that time averaged SSTs remained close to observed climatological values in the control simulations. However SSTs were allowed to vary freely

in response to natural and forced variations. The selection of parameter perturbations was designed to sample uncertainties in a wide range of processes without making a priori assumptions about the relative importance of different climate change feedbacks⁴. The perturbations affected large-scale cloud and precipitation, convection, radiation, dynamics, boundary layer transports, land surface processes and sea ice. Parameters were perturbed either by changing a logical switch or by setting a coefficient or threshold to a minimum, intermediate or maximum value specified by experts, one of these (often, but not always, the intermediate value) being that used in STD. See Supplementary Information for details.

Weighting the predictions of ensemble members

We seek to weight model predictions of climate sensitivity according to the likelihood that the simulation of present-day climate is consistent with observations. The probability that a simulated variable m belongs to a population of observations of mean o and standard deviation σ is proportional to $\exp\{-0.5(m-o)^2/\sigma^2\}$, assuming gaussian statistics. In principle, the likelihood of the model can be obtained by calculating the joint probability of all model variables, taking into account their covariances and allowing for errors in the verifying observations. In practice, the required error statistics were not available, because our model versions were not run long enough to estimate their covariance matrices and observational errors are not known for most of the variables included in the CPI. We therefore make simplifying assumptions in order to obtain a likelihood-based weight. Our choice is $\exp(-0.5\text{CPI}^2)$, which represents an estimate of likelihood obtained by normalizing the error variance in simulated climate by the variance of simulated interannual variations and then averaging the normalized error variance over a wide range of climate variables. Our decision to weight each component equally when forming the CPI represents an a priori assumption that changes in climate sensitivity are equally affected by all model variables.

Production of probability distributions for climate sensitivity

We obtained statistical predictions of CPI and climate sensitivity for 4×10^6 random combinations of multiple parameter perturbations generated by assuming a uniform prior for each parameter within the range specified by experts. Predictions of climate sensitivity (ΔT) were made in terms of the feedback strength λ , defined as $\lambda = \Delta Q/\Delta T$, ΔQ being the radiative forcing due to doubling CO₂. We predicted values of CPI and λ by assuming that the effects of individual parameter perturbations on present-day climate fields and feedback strength can be interpolated linearly between the values sampled in our PPE, and that the effects of individual parameter perturbations combine linearly and independently.

The predictions of λ are sensitive to λ_{std} , the feedback strength found in STD, since all perturbations are calculated relative to this value. We therefore repeated the calculation for 21 values of λ_{std} sampling at equal intervals the $\pm$ two standard deviation uncertainty range of 600 yr mean values (1.069–1.088 W m⁻² K⁻¹) estimated from 600 yr control and $2 \times \text{CO}_2$ integrations of STD. For each of these $21 \times 4 \times 10^6$ predictions, we calculated an

associated error as the sum of terms arising from nonlinear interactions between parameter perturbations and noise (natural variability) in the model simulations used to construct the predictions. Both error terms were assumed to be independent of location in parameter space (and hence climate sensitivity). The first term was estimated by verifying our statistical predictions against simulations made with 13 model versions containing multiple parameter perturbations and simulating climate sensitivities in the range 3.1–4.9 °C. The second term was estimated from the long STD experiment (See Supplementary Information). Each of our $21 \times 4 \times 10^6$ predictions of λ was then expressed as a gaussian distribution accounting for its expected error. A PDF of feedback strength was derived by combining the resulting $21 \times 4 \times 10^6$ distributions, each weighted according to the probability of the relevant value of λ_{std} . This was converted into a PDF of climate sensitivity using $\Delta T = \Delta Q/\lambda$, giving the blue PDF in Fig. 3. The red PDF was derived in the same manner, except that a further weighting of $\exp(-0.5\text{CPI}^2)$ was applied to each of the gaussian distributions of λ . Results from our 13 verifying multiple perturbation experiments showed that our statistical predictions of CPI were close to the simulated values and that the predictions of λ carried a standard error of about $0.15 \text{ W m}^{-2} \text{ K}^{-1}$, arising mainly from the nonlinear effects of combining parameter perturbations²⁹.

Received 29 March; accepted 16 June 2004; doi:10.1038/nature02771.

- Gordon, C. *et al.* The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim. Dyn.* **16**, 147–168 (2000).
- Giorgi, F. & Francisco, R. Evaluating uncertainties in the prediction of regional climate change. *Geophys. Res. Lett.* **27**, 1295–1298 (2000).
- Allen, M. R. & Ingram, W. J. Constraints on future changes in climate and the hydrological cycle. *Nature* **419**, 224–232 (2002).
- Cubasch, U. *et al.* in *Climate Change 2001, The Science of Climate Change* Ch. 9 (eds Houghton, J. T. *et al.*) 527–582 (Cambridge Univ. Press, Cambridge, 2001).
- Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Delworth, T. L. Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **417**, 617–620 (2000).
- Mearns, L. O. *et al.* in *Climate Change 2001, The Science of Climate Change* Ch. 13 (eds Houghton, J. T. *et al.*) 739–768 (Cambridge Univ. Press, Cambridge, 2001).
- Meehl, G. A. *et al.* Trends in extreme weather and climate events: issues related to modelling extremes in projections of future climate change. *Bull. Am. Meteorol. Soc.* **81**, 427–436 (2000).
- Stocker, T. F. *et al.* in *Climate Change 2001, The Science of Climate Change* Ch. 7 (eds Houghton, J. T. *et al.*) 417–470 (Cambridge Univ. Press, Cambridge, 2001).
- Randall, D. A. & Wielicki, B. A. Measurements, models and hypotheses in the atmospheric sciences. *Bull. Am. Meteorol. Soc.* **78**, 399–406 (1997).
- Wigley, T. M. L. & Raper, S. C. B. Interpretation of high projections for global-mean warming. *Science* **293**, 451–454 (2001).
- Palmer, T. N. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q.J.R. Meteorol. Soc.* **127**, 279–303 (2001).
- Pope, V. D., Gallani, M., Rowntree, P. R. & Stratton, R. A. The impact of new physical parameterisations in the Hadley Centre climate model—HadAM3. *Clim. Dyn.* **16**, 123–146 (2000).
- Keen, A. B. & Murphy, J. M. Influence of natural variability and the cold start problem on the simulated transient response to increasing CO₂. *Clim. Dyn.* **13**, 847–864 (1997).
- Palmer, T. N. & Räisänen, J. Quantifying the risk of extreme seasonal precipitation events in a changing climate. *Nature* **415**, 514–517 (2002).
- Leroy, S. S. Detecting climate signals: Some Bayesian aspects. *J. Clim.* **11**, 640–651 (1998).
- Stott, P. A. & Kettleborough, J. A. Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726 (2002).
- Giorgi, F. & Mearns, L. O. Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the “Reliability Ensemble Averaging” (REA) method. *J. Clim.* **15**, 1141–1158 (2002).
- Giorgi, F. & Mearns, L. O. Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.* **30**, 1629 (2003).
- Andronova, N. G. & Schlesinger, M. E. Objective estimation of the probability density function for climate sensitivity. *J. Geophys. Res.* **106**, 22605–22612 (2001).
- Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, M. D. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117 (2002).
- Knutti, R., Stocker, T. F., Joos, F. & Plattner, G. K. Constraints on radiative forcing and future climate change from observations and climate model ensembles. *Nature* **416**, 719–723 (2002).
- Gregory, J. M., Stouffer, R. J., Raper, S. C. B., Stott, P. A. & Rayner, N. A. An observationally based estimate of the climate sensitivity. *J. Clim.* **15**, 3117–3121 (2002).
- Anderson, T. L. *et al.* Climate forcing by aerosols—a hazy picture. *Science* **300**, 110–111 (2003).
- Barnett, T. P., Pierce, D. W. & Schnur, R. Detection of anthropogenic climate change in the world's oceans. *Science* **292**, 270–274 (2001).
- Boer, G. J. & Yu, B. Dynamical aspects of climate sensitivity. *Geophys. Res. Lett.* **30**, 1135 (2003).
- Stouffer, R. J. & Manabe, S. Response of a coupled ocean-atmosphere model to increasing atmospheric carbon dioxide: sensitivity to the rate of increase. *J. Clim.* **12**, 2224–2237 (1999).
- Watterson, I. G. Interpretation of simulated global warming using a simple model. *J. Clim.* **13**, 202–215 (2000).
- Senior, C. A. & Mitchell, J. F. B. The time-dependence of climate sensitivity. *Geophys. Res. Lett.* **27**, 2685–2689 (2000).
- Stainforth, D. A. *et al.* Evaluating uncertainty in the climate response to changing levels of greenhouse gases. *Nature* (submitted).
- Allen, M. R. & Stainforth, D. A. Towards objective probabilistic climate forecasting. *Nature* **419**, 228 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. Thorpe and M. Allen for encouraging the development of this project. We also acknowledge many Hadley Centre colleagues for their advice on model parameters and their uncertainty ranges, and for comments on earlier versions of the manuscript. This work was supported by the UK Department of the Environment, Food and Rural Affairs. D.A.S. was funded by the NERC COAPEC thematic programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.M. (james.murphy@metoffice.com).

Gigantism and comparative life-history parameters of tyrannosaurid dinosaurs

Gregory M. Erickson^{1,2}, Peter J. Makovicky³, Philip J. Currie⁴, Mark A. Norell², Scott A. Yerby⁵ & Christopher A. Brochu⁶

¹Department of Biological Science, Florida State University, Tallahassee, Florida 32306-1100, USA

²Division of Paleontology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

³Department of Geology, The Field Museum, 1400 S. Lake Shore Drive, Chicago, Illinois 60605, USA

⁴Royal Tyrrell Museum of Palaeontology, Drumheller, Alberta T0J 0Y0, Canada

⁵Department of Biomechanical Engineering, Stanford University, Stanford, California 94305, USA

⁶Department of Geoscience, University of Iowa, Iowa City, Iowa 52242, USA

How evolutionary changes in body size are brought about by variance in developmental timing and/or growth rates (also known as heterochrony) is a topic of considerable interest in evolutionary biology¹. In particular, extreme size change leading to gigantism occurred within the dinosaurs on multiple occasions². Whether this change was brought about by accelerated growth, delayed maturity or a combination of both processes is unknown. A better understanding of relationships between non-avian dinosaur groups and the newfound capacity to reconstruct their growth curves make it possible to address these questions quantitatively³. Here we study growth patterns within the Tyrannosauridae, the best known group of large carnivorous dinosaurs, and determine the developmental means by which *Tyrannosaurus rex*, weighing 5,000 kg and more, grew to be one of the most enormous terrestrial carnivorous animals ever. *T. rex* had a maximal growth rate of 2.1 kg d^{-1} , reached skeletal maturity in two decades and lived for up to 28 years. *T. rex*'s great stature was primarily attained by accelerating growth rates beyond that of its closest relatives.

Stemming from more than a century of investigation, considerable understanding of tyrannosaurid osteology⁴, myology⁵, neurology⁶, behaviour^{7,8}, physiology^{3,9}, physical capabilities^{10,11} and phylogeny^{12,13} have been gained. Lacking are empirical data on tyrannosaurid life history such as growth rates, longevity and somatic maturity (adult size) from which the developmental possibilities for how *T. rex* attained gigantism can be formally tested.

Recent advances in techniques for determining the ages at death of dinosaurs by using skeletal growth line counts^{3,14}, coupled with developmental size estimates³, make quantitative growth-curve reconstructions for dinosaurs feasible. These methods have been used to study growth rates in two small theropods, a small and a large ornithischian and a medium-sized and a gigantic sauropodomorph³. These data were used to derive a regression of body mass against growth rate and to generalize broadly about non-avian

dinosaur growth³. However, because of the phylogenetically disparate nature of these data (that is, none are close outgroups to one another) it has not been possible to use them to infer how specific cases of size change occurred within dinosaurian sub-clades such as the Tyrannosauridae. Such an understanding requires multi-species sampling at low taxonomic levels (that is, among closely related species) and access to growth series spanning juvenile through adult stages, a rarity among extinct dinosaurs¹⁵. Furthermore, it requires the capacity to account for growth line losses due to medullar cavity hollowing and cortical remodelling¹⁶, two processes that are pervasive in the major weight-bearing bones from large theropods such as tyrannosaurids.

The sampling problem has been overcome in North American tyrannosaurids. A flurry of recent discoveries has greatly increased the number of substantially complete specimens representing various growth stages available for study. For example, more than 30 *T. rex* specimens are known^{4,17}, compared with only 11 reported in 1993 (ref. 18; see Supplementary Information). Recent work has broadened the developmental representation of these animals by showing that several purported 'dwarf' tyrannosaur species are juveniles of larger, previously recognized forms such as *T. rex*^{12,13,19,20}. Finally, preliminary analyses for this research revealed that several non-weight-bearing bones in tyrannosaurids (for example pubes, fibulae, ribs, gastralia and postorbitals) did not develop hollow medullar cavities and showed negligible intracortical remodelling during their entire life history (Fig. 1). Like major long bones, these elements are effective for assessing longevity in living reptiles (Fig. 1)^{21,22} and hence provide a viable alternative method for determining the age at death of extinct reptiles such as tyrannosaurids.

Here we exploit these findings to determine the pattern of growth in *T. rex* and three of its close tyrannosaurid relatives. We then use character optimization methods²³ to infer how *T. rex* attained giant proportions among tyrannosaurids. Finally, this new evidence is used to further our understanding of tyrannosaurid biology. In performing these analyses, we sampled several amedullar bones from adolescent, juvenile, sub-adult and adult representatives of the North American Late Cretaceous tyrannosaurids *Albertosaurus sarcophagus*, *Gorgosaurus libratus*, *Daspletosaurus torosus* and *T. rex*. Longevity in each of the 20 specimens was assessed from line counts in histological sections by using polarizing, dissecting and reflected-light microscopy (Fig. 1)^{3,14}. Conservative estimates of body mass (see Supplementary Information) were made by using femoral circumference measures²⁴. Longevity and size data were plotted and least-squares regression was used to determine the first empirical growth curves for tyrannosaurids³. The length and timing of the various developmental stages and the maximal growth rates for each taxon were compared²⁵. The results were examined in an evolutionary context²³ by using two competing phylogenetic hypotheses for the Tyrannosauridae^{12,13}.

Sampled longevity for *T. rex* ranged from 2 to 28 years and corresponding body mass estimates ranged from 29.9 to 5,654 kg (Table 1). The transition to somatic maturity in this taxon seems to have begun at about 18.5 years of age (Fig. 2). At least one individual (exemplified by FMNH (The Field Museum) PR 2081), showed evidence for prolonged senescence in the form of conspicuously narrow pericortical growth-line spacing (Fig. 1). Maximal growth rates in *T. rex* were 2.07 kg d^{-1} and such exponential rates were maintained for about 4 years (Fig. 2). The longevity estimates for *T. rex* outgroups ranged from 2 to 24 years and corresponding body sizes spanned from 50.3 to 1,791 kg (Table 1). Somatic maturity occurred at between 14 and 16 years in these taxa (Fig. 2). Like *T. rex*, at least some exceptionally large individuals of *A. sarcophagus* and *D. torosus* showed narrow pericortical growth-line spacing indicative of the onset of senescence. The maximal growth rates for the three smaller tyrannosaurid taxa ranged from 0.31 to 0.48 kg d^{-1} ; such exponential stage rates were also maintained for about 4 years (Fig. 2). Optimization of growth rates onto the two current

phylogenetic hypotheses of tyrannosaurid relationships suggests that a 1.5-fold acceleration in maximal growth rate might diagnose Tyrannosaurinae (the clade comprising *Daspletosaurus* and *Tyrannosaurus*^{13,19}, Fig. 2). A second substantial increase in growth rate optimizes as a physiological autapomorphy of *Tyrannosaurus* irrespective of phylogenetic hypothesis and optimization criterion.

T. rex is notable for its great size, which is at least 15-fold greater than the largest living terrestrial carnivorous animals today and second only to *Giganotosaurus*²⁶ among theropod dinosaurs. How did it attain such great proportions within the Tyrannosauridae? From the two competing hypotheses of tyrannosaurid phylogeny it is most parsimonious to conclude that *T. rex* acquired the majority of its giant proportions after diverging from the common ancestor of itself and *D. torosus*, a species with an optimized body mass of about 1,800 kg. Direct comparison between the tyrannosaurid growth curves shows that the transition to the exponential and stationary phases of development occurred about 2–4 years later in *T. rex* (Fig. 2). However, such temporal post-displacement had little to do with the evolution of its gigantism because the exponential

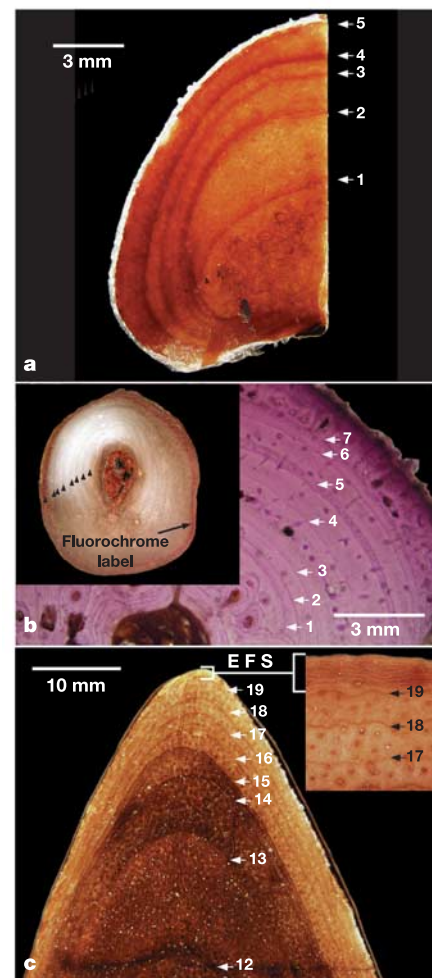


Figure 1 Growth-line counts in tyrannosaurids and reptiles of known ages. **a**, Thin-sectioned *Gorgosaurus* fibula (FMNH PR 2211). The growth record in this element is complete and shows five growth lines (arrows), indicating that it died early in its sixth year of life. **b**, Haematoxylin-stained fibula from an 8-year-old alligator (*Alligator mississippiensis*) showing the expected seven growth lines. Inset, gastralia (unstained) from 9-year-old *A. mississippiensis* showing the expected eight growth lines and the chemical label used to verify the periodicity of ring formation. **c**, *Tyrannosaurus* rib (FMNH PR 2081) showing the 15th to 19th growth lines. Inset, external fundamental system¹⁶ with nine tightly spaced growth lines, indicating late-adulthood senescence and growth-rate attenuation.

Table 1 Specimens, size measures, and longevity data

Taxon	Specimen no.	Elements examined	Femoral length (cm)	Body mass (kg)	Longevity (years); no. of growth lines in EFS
<i>Tyrannosaurus rex</i>	FMNH PR 2081	G, R, F, OLB	134.5	5654	28; 9
	RTMP 81.12.1	R, OLB	128.4 est.	5040	22
	RTMP 81.6.1	G, R, OLB	120	3230	18
	ICM 2001.90.1	G, R	116.8 est.	2984	16
	LACM 23845	P, OLB	98.9	1807	14
	AMNH 30564	G, R, F	98	1761	15
	LACM 28471	R, C	25.2 est.	29.9	2
<i>Gorgosaurus libratus</i>	RTMP 94.12.602	R, G, OLB	91.6	1105	18
	RTMP 73.30.1	F, OLB	80.4 est.	747	14
	RTMP 99.33.1	G, F, OLB	75	607	14
	RTMP 86.144.1	R, F, OLB	54.2	229	7
	FMNH PR 2211	G, R, F, OLB	44.5	127	5
<i>Albertosaurus sarcophagus</i>	RTMP 81.10.1	G, R, F, OLB	89.5	1142	24; 2
	AMNH 5432	P, F	99.3 est.	1282	22
	USNM 12814/AMNH 5428	R	86.0 est.	1013	18
	RTMP 86.64.01	R, F, OLB	78.2 est.	762	15
	RTMP 2002.45.46	F, OLB	31.6 est.	50.3	2
<i>Daspletosaurus torosus</i>	FMNH PR 308	G, R	96.0 est.	1791	21; 5
	AMNH 5438	R	84.1 est.	1518	17
	RTMP 94.143.1	G, OLB	62.6 est.	496	10

FMNH, The Field Museum; RTMP, Royal Tyrrell Museum of Palaeontology; ICM, Indianapolis Children's Museum; LACM, Los Angeles County Museum; AMNH, American Museum of Natural History; USNM, United States National Museum. R, rib; G, gastralia; F, fibula; P, pubis; C, dermal skull bones; OLB, other long bones; est., estimated; EFS, external fundamental system¹⁶.

stage, during which most body size is accrued²⁵, was not extended beyond the ancestral, 4-year condition observed in other tyrannosaurids. Rather, the key developmental modification that propelled *T. rex* to giant proportions was primarily through evolutionary acceleration in the exponential stage growth rate and the transition zones bounding it. This is reflected in the regions of maximal slope on the growth curves depicted in Fig. 2 and holds true regardless of which evolutionary hypothesis is correct and how the maximum growth rates are optimized. Notably, this method of attaining gigantism contrasts with that in the largest crocodylians and lizards, where ancestral growth rates were retained and the exponential stages lengthened²⁷. How other dinosaurs attained gigantism within their respective sub-clades will serve as an interesting line of inquiry in the future. Does the same pattern of acceleratory growth seen here characterize the means by which all or most members of the Dinosauria attained great size?

Besides revealing how the evolution of *T. rex* gigantism was obtained, the data garnered here provide for a more comprehensive understanding of tyrannosaurid biology. For instance the presence of thin, tightly packed growth lines late in development (Fig. 1) shows that these animals, like nearly all (if not all) dinosaurs, had determinate growth^{3,14}. They would not have gained an appreciably greater size than the largest specimens studied here and could spend nearly 30% of their lives as full-grown adults (Fig. 2). In addition, the maximal growth rates for these tyrannosaurid species are only 33–52% of the rates expected for non-avian dinosaurs of their size when compared with the more broadly sampled data of Erickson *et al.*³. This provides the first evidence of its kind pointing to major differences in whole body growth rates among a non-avian dinosaur sub-clade. Such findings are not unexpected because similar patterns (for example primates within Eutheria) occur within living vertebrate groups²⁸. Our findings also have a bearing on the

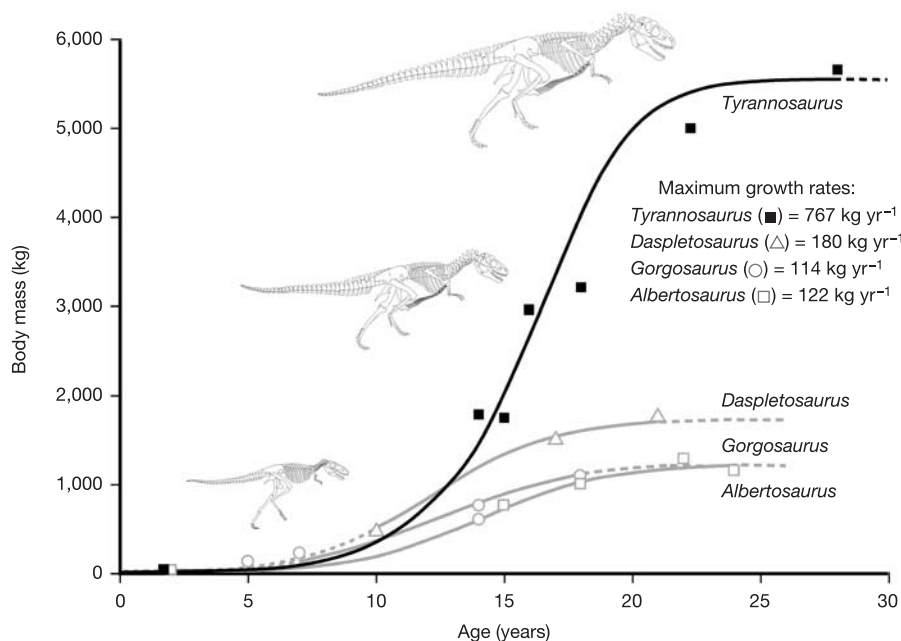


Figure 2 Logistic growth curves for *Tyrannosaurus* and three related tyrannosaurids. Note that the exponential stages (the regions of maximal slope) are similar in duration but differ in slope (that is, growth rates). Regression equations (mass in kg, age in years) are as follows: *T. rex*, mass = {5,551/[1 + e^{-0.57(age - 16.1)}]} + 5, $r^2 = 0.953$; *D. torosus*,

mass = {1,728/[1 + e^{-0.44(age - 12.1)}]} + 5, $r^2 = 0.992$; *G. libratus*, mass = {1,234/[1 + e^{-0.38(age - 12.4)}]} + 5, $r^2 = 0.950$; *A. sarcophagus*, mass = {1,218/[1 + e^{-0.43(age - 14.1)}]} + 5, $r^2 = 0.985$.

biomechanical capacities of tyrannosaurids. *T. rex*'s capacity for 'fast running' was biomechanically infeasible after a body mass of about 1,000 kg was attained¹¹. This corresponds to a juvenile-sized animal just 13 years of age on the basis of our longevity data and conservative estimates of body mass (Fig. 2). If we assume that the same relationship held true for the smaller tyrannosaurid species studied here, such locomotory limitations would not have emerged until these animals were much closer to adult size (Fig. 2). Finally, a glimpse into the potential population age structure for a dinosaur is also afforded from these data. Currie⁷ has described a catastrophic death assemblage consisting of eight or nine *A. sarcophagus* specimens thought to represent an entire pack or a subset of one. On the basis of femoral lengths, the age and developmental stage of each animal can now be estimated. The group seems to have consisted of two or three older adults ~21 or more years of age, one ~17-year-old young adult, four ~12–17-year-old sub-adults that were undergoing exponential stage growth at the time of death, and one ~10-year-old juvenile that was beginning the transition to exponential stage growth. A reopening of the site has revealed at least one more specimen (RTMP (Royal Tyrrell Museum of Palaeontology) 2002.45.46) shown here to be only 2 years old (Table 1). This indicates that *A. sarcophagus* groups, whether temporary or permanent, might have been composed of individuals spanning the age spectrum from adolescents to very old, senescent adults, a finding consistent with trackway evidence for other theropod dinosaurs⁷. □

Methods

Assessments of tyrannosaurid longevity

Growth lines have been shown to form throughout the cranial and post-cranial skeletons of living tetrapods^{16,21,22,27}, and their annual formative rhythms have been shown in lepidosaurian and archosaurian (that is crocodilian) outgroups to dinosaurs^{21,22,27}. This indicates that similar annual genesis might have occurred in non-avian dinosaurs such as tyrannosaurids³. Here we used fibulae, pubes, gastralia, ribs and cranial bones and a few selected long-bone elements in which medullary cavity expansion and remodelling was not pervasive. Bones such as these have been shown to have excellent efficacy for assessing longevity in snakes²¹ and lizards²², and our own multi-element histological studies on lizards and crocodilians of known ages from the Florida Museum of Natural History, Gainesville, and the Florida Fish and Wildlife Conservation Commission, Gainesville, confirm this (Fig. 1). In cases where some of the earliest-forming tyrannosaur growth rings were remodelled, losses were accounted for by examining younger individuals in which the rings were still present or through back-calculation methods³. Inter-elemental comparisons revealed longevity estimates to within 1 year in all cases, the higher values of which were used in the longevity assessments.

Body size estimates

To permit comparisons of growth rates between tyrannosaurids and living taxa and to negate the effects of shape differences, the data were standardized to body mass³ by using a femoral circumference/body mass regression equation²⁴. This method produces conservative estimates of body mass²⁹ (see Supplementary Information). Subadult sizes were determined with the Developmental Mass Extrapolation femoral scaling principle³. In cases in which the femora were not preserved, the tyrannosaurid long-bone regression equations of Currie³⁰ were used to estimate the missing measurements.

Reconstruction of growth curves

Age and mass data were contrasted to reveal growth curves for the four tyrannosaurid species (Fig. 2). All vertebrates show logistic (S-shaped) growth patterns such as these during post-parturition/hatching development^{3,25}. Hence a logistic equation and least-squares regression analysis were used to describe the relationship between the data. The terminology used to refer to the various stages of development is as follows²⁵. The initial, slow growth phase is referred to as the lag stage. This is followed by the linear exponential stage during which maximal slope is observed and maximal growth rates occur. Finally, growth reaches a plateau during the stationary stage. Here we refer to animals in the lag stage as adolescents, those in transition to the exponential stage as juveniles, those in the exponential stage as sub-adults, those in transition to the stationary stage as young adults, and finally animals in the stationary stage as senescent adults. Here, adulthood refers to the onset of somatic maturity whereby adult size is first attained. Is not known when sexual maturity for non-avian dinosaurs occurred, because gender has not been established with certainty for any taxon.

Character evolution analysis

Maximum growth rates were optimized as a continuous character onto two competing phylogenies for the Tyrannosauridae^{12,13}. The topologies as they pertain to the taxa sampled are (*Albertosaurus*, *Gorgosaurus* (*Daspletosaurus*, *Tyrannosaurus*))¹² and (*Albertosaurus*, *Gorgosaurus*) (*Daspletosaurus*, *Tyrannosaurus*)¹³. Because no basal

tyrannosaurids have been sampled, the basal theropod *Syntarsus*³ was used to root the trees. Both linear and squared-changes optimizations were applied²³. Both optimality criteria and all possible topologies suggest that a large increase in growth rate is diagnostic of the *Daspletosaurus*–*Tyrannosaurus* clade (Tyrannosaurinae)¹³, and a second large increment in growth rate optimizes as unique to *T. rex*. The actual magnitude of the growth rate change reconstructed at ancestral nodes differs with topology and more drastically with the optimization method. Linear parsimony yields a punctuated pattern with higher changes at individual nodes, whereas squared-changes parsimony forces a 'smoother' distribution on the data but also incurs some counterintuitive deceleration in growth for the slower-growing basal taxa.

Received 3 March; accepted 26 May 2004; doi:10.1038/nature02699.

- Gould, S. J. *Ontogeny and Phylogeny* (Harvard Univ. Press, Cambridge, Massachusetts, 1977).
- Carrano, M. T. in *Amniote Paleobiology: Perspectives on the Evolution of Mammals, Birds, and Reptiles* (eds Carrano, M. T., Bloch, R. W., Gaudin, T. & Wible, J.) (Univ. Chicago Press, Chicago, in the press).
- Erickson, G. M., Curry-Rogers, K. & Yerby, S. Dinosaur growth patterns and rapid avian growth rates. *Nature* **412**, 429–433 (2001).
- Brochu, C. A. Osteology of *Tyrannosaurus rex*: Insights from a nearly complete skeleton and high-resolution computed tomographic analysis of the skull. *J. Vert. Paleontol. Mem.* **22**(7), 1–138 (2003).
- Carrano, M. T. & Hutchinson, J. R. Pelvic and hindlimb musculature of *Tyrannosaurus rex* (Dinosauria: Theropoda). *J. Morphol.* **253**, 207–228 (2002).
- Brochu, C. A. A digitally rendered endocast for *Tyrannosaurus rex*. *J. Vert. Paleontol.* **20**, 1–6 (2003).
- Currie, P. J. Possible evidence of gregarious behavior in tyrannosaurids. *GAIA* **15**, 271–277 (1998).
- Erickson, G. M. Breathing life into *Tyrannosaurus rex*. *Scient. Am.* **9**, 42–49 (1999).
- Chin, K., Tokaryk, T. T., Erickson, G. M. & Calk, L. C. A king-sized theropod coprolite. *Nature* **393**, 680–682 (1998).
- Erickson, G. M. *et al.* Bite-force estimation for *Tyrannosaurus rex* from tooth-marked bones. *Nature* **382**, 706–708 (1996).
- Hutchinson, J. & Garcia, M. *Tyrannosaurus* was not a fast runner. *Nature* **415**, 1018–1021 (2002).
- Holtz, T. R. in *Mesozoic Vertebrate Life* (eds Tanke, D. & Carpenter, K.) 64–83 (Indiana Univ. Press, Bloomington, 2001).
- Currie, P. J. Cranial anatomy of tyrannosaurid dinosaurs from the Late Cretaceous of Alberta, Canada. *Acta Palaeontol. Pol.* **48**, 191–226 (2003).
- Chinsamy, A. Physiological implications of the bone histology of *Syntarsus rhodesiensis* (Saurischia: Theropoda). *Palaeontol. Afr.* **27**, 77–82 (1993).
- Weishampel, D. B. & Horner, J. R. in *Dinosaur Eggs and Babies* (eds Carpenter, K., Hirsch, K. F. & Horner, J. R.) 229–243 (Cambridge Univ. Press, Cambridge, 1994).
- Francillon-Viellet, H. *et al.* in *Biomimeticization: Patterns and Evolutionary Trends* (ed. Carter, J. G.) 471–530 (Van Nostrand Reinhold, New York, 1990).
- Larson, P. & Donnan, K. *Rex Appeal* (Invisible Cities Press, Montpelier, 2002).
- Horner, J. R. & Lessem, D. *The Complete T. rex* (Simon & Schuster, New York, 1993).
- Currie, P. J., Hurum, J. H. & Sabath, K. Skull structure and evolution in tyrannosaurid dinosaurs. *Acta Palaeontol. Pol.* **48**, 227–234 (2003).
- Carr, T. D. Craniofacial ontogeny in Tyrannosauridae (Dinosauria, Coelurosauria). *J. Vert. Paleontol.* **19**, 497–520 (1999).
- Castanet, J. & Naulleau, G. Données expérimentales sur la valeur des marques squelettiques comme indicateur de l'âge chez *Vipera aspis* (L.) (Ophidia, Viperidae). *Zool. Scripta* **3**, 201–208 (1974).
- de Buffrenil, V. & Castanet, J. Age estimation by skeletochronology in the Nile monitor (*Varanus niloticus*), a highly exploited species. *J. Herpetol.* **34**, 414–424 (2000).
- Maddison, W. P. & Maddison, D. R. *MacClade Version 3: Analysis of Phylogeny and Character Evolution* (Sinauer Associates, Sunderland, MA, 1992).
- Anderson, J. F., Hall-Martin, A. & Russell, D. A. Long bone circumference and weight in mammals, birds, and dinosaurs. *J. Zool. A* **207**, 53–61 (1985).
- Sussman, M. *Growth and Development* (Prentice-Hall, Englewood Cliffs, New Jersey, 1964).
- Carr, R. A. & Salgado, L. A new giant carnivorous dinosaur from the Cretaceous of Patagonia. *Nature* **377**, 224–226 (1995).
- Erickson, G. M., de Ricqlès, A., de Buffrenil, V., Molnar, R. E. & Bayless, M. A. Vermiform bones and the evolution of gigantism in *Megalania*—how a reptilian fox became a lion. *J. Vert. Paleontol.* **23**, 966–970 (2003).
- Case, T. E. On the evolution and adaptive significance of post-natal growth rates in the terrestrial vertebrates. *Q. Rev. Biol.* **53**, 243–282 (1978).
- Seebacher, F. A new method to calculate allometric length-mass relationships of dinosaurs. *J. Vert. Paleontol.* **21**, 51–60 (2001).
- Currie, P. J. Allometric growth in tyrannosaurids (Dinosauria: Theropoda) from the Upper Cretaceous of North America and Asia. *Can. J. Earth Sci.* **40**, 651–665 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank O. Rieppel and B. Simpson (The Field Museum), L. Chiappe (Los Angeles County Museum of Natural History), P. Larson and N. Larson (Black Hills Institute), B. Stein and M. Triebold (Triebold Paleontology), D. Evans (Indianapolis Children's Museum), C. Mehling (American Museum of Natural History), S. Williams, M. Henderson and L. Cranford (Burpee Museum of Natural History), J. Gardner, D. Tanke and D. Brinkman (Royal Tyrrell Museum of Palaeontology), T. Carr (University of Toronto), F. W. King and K. Krysko (Florida Museum of Natural History), K. Womble (Florida State University), M. Bayless (Berkeley, California) and A. Woodward (Florida Fish and Wildlife Conservation Commission) for assistance with this research. The NSF and the College of Arts and Sciences of Florida State University generously funded this project.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.M.E. (gerickson@bio.fsu.edu).

Eocene evolution of whale hearing

Sirpa Nummela¹, J. G. M. Thewissen¹, Sunil Bajpai², S. Taseer Hussain³ & Kishor Kumar⁴

¹Department of Anatomy, Northeastern Ohio Universities College of Medicine, Rootstown, Ohio 44272, USA

²Department of Earth Sciences, Indian Institute of Technology, Roorkee 227 667, Uttaranchal, India

³Department of Anatomy, Howard University, College of Medicine, Washington DC 20059, USA

⁴Wadia Institute of Himalayan Geology, Dehradun 248 001, India

The origin of whales (order Cetacea) is one of the best-documented examples of macroevolutionary change in vertebrates^{1–3}. As the earliest whales became obligately marine, all of their organ systems adapted to the new environment. The fossil record indicates that this evolutionary transition took less than 15 million years, and that different organ systems followed different evolutionary trajectories. Here we document the evolutionary changes that took place in the sound transmission mechanism of the outer and middle ear in early whales. Sound transmission mechanisms change early on in whale evolution and pass through a stage (in pakicetids) in which hearing in both air and water is unsophisticated. This intermediate stage is soon abandoned and is replaced (in remingtonocetids and protocetids) by a sound transmission mechanism similar to that in modern toothed whales. The mechanism of these fossil whales lacks sophistication, and still retains some of the key elements that land mammals use to hear airborne sound.

The cetacean sound transmission mechanism has long been known to be near-modern less than 10 million years after the origin

of the order Cetacea^{4–6}. Fossils documenting transitional ear morphologies were suggestive of pronounced change^{7,8}, but could not be interpreted in a functional context because of their scarcity and because the function of modern cetacean sound transmission was not understood until recently^{9–11}.

Our newly discovered cetacean fossils (Fig. 1) allow us to present an integrated interpretation of evolving sound transmission mechanisms as whales took to the water. Our specimens represent four groups of early whales³. These include pakicetids (*Ichthyolestes*, *Nalacetus* and *Pakicetus*), the basal cetacean clade known from 50-million-year-old deposits in Pakistan. They also include remingtonocetids (*Andrewsiphium* and *Remingtonocetus*) and protocetids (*Indocetus*), which lived approximately 43–46 million years ago in India, and represent taxa that are successively more closely related to modern cetaceans. Finally, we studied previously described material for basilosaurids⁵ (*Basilosaurus* and *Zygorhiza*) from North America, which lived from approximately 35 to 40 million years ago. We did not include ambulocetid whales in this study because no ear ossicles are known for this family.

The ear of the earliest whales, pakicetids, functions in a similar way to that of land mammals when receiving airborne sound (Fig. 2)^{8,12,13}. Sounds are transmitted through the air-filled external auditory meatus and cause the tympanic membrane to vibrate¹⁴. Its vibrations are passed on through a chain of small bones (middle ear ossicles) located in the air-filled middle ear cavity. The last of these ossicles, the stapes, acts as a piston that causes vibrations in the inner ear fluid. Although adequate in air, this mechanism fails underwater where the external auditory meatus fills with water, creating a pressure differential across the tympanic membrane that greatly diminishes its ability to transmit sound. In addition, underwater sound will reach the pakicetid ear by passing through the head tissues. This mode of hearing is called bone conduction¹⁵ and does

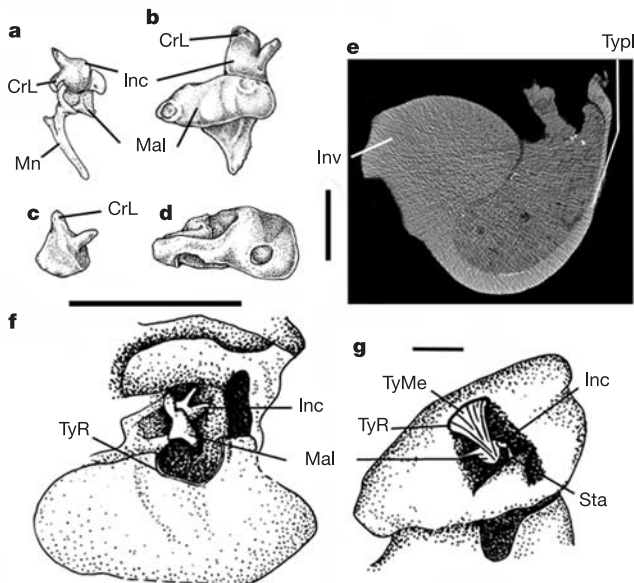


Figure 1 Sound transmission elements in cetaceans and an artiodactyl. **a, b**, Left malleus and incus of a pig (**a**) and a bottlenose dolphin (**b**). **c**, Left incus of *Pakicetus* (H-GSP 91035). **d**, Right malleus of *Indocetus* (LUVF 11034). Ossicles are drawn perpendicular to the plane of the lever arms. **e**, Computerised tomography (CT) scan of a section through the left tympanic of *Remingtonocetus* (RUSB 2914). **f, g**, Left ear of *Remingtonocetus* (RUSB 2828) in rostro-lateral (**f**) and ventral (**g**) view, with the three ossicles *in situ*, the tympanic ring (TyR) position projected and the tympanic membrane (TyMe) position inferred. Scale bar for **f** and **g** is beside **g**. Abbreviations: CrL, crus longum; Inc, incus; Inv, involucrum; Mal, malleus; Mn, manubrium; Sta, stapes; TyPl, tympanic plate. Scale bars, 10 mm.

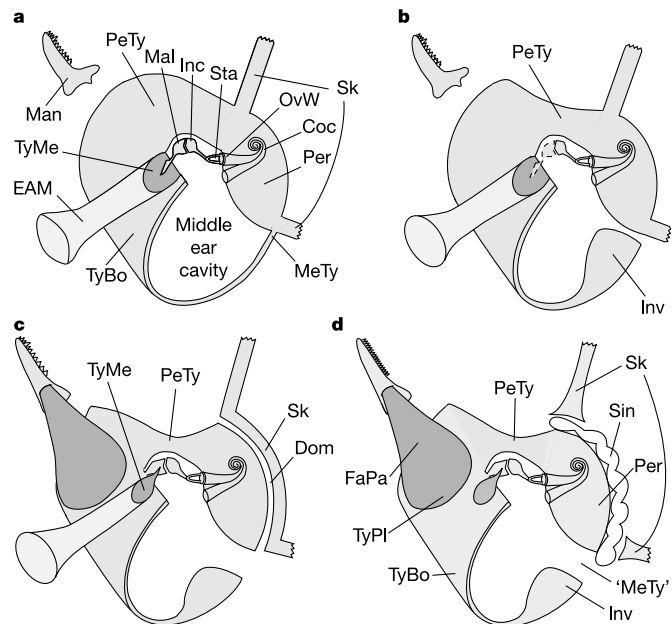


Figure 2 Sound transmission mechanisms in land mammals and whales. Diagram of the ear in a generalized land mammal (**a**), a pakicetid (**b**), a remingtonocetid/protocetid (**c**) and a modern odontocete (**d**). Abbreviations: Coc, cochlea; Dom, dome-shaped depression for periotic; EAM, external auditory meatus; FaPa, fat pad; Inc, incus; Inv, involucrum; Mal, malleus; Man, mandible; MeTy, medial synostosis between periotic and tympanic bone, in cetaceans this synostosis is absent and is homologous to a gap between these bones ('MeTy'); OvW, oval window; Per, periotic bone; PeTy, joint between periotic and tympanic; Sin, air-filled sinuses; Sk, skull; Sta, stapes; TyBo, tympanic bone; TyMe, tympanic membrane; TyPl, tympanic plate.

not allow for directional hearing. In air, bone conduction is nearly absent because the air–body interface reflects most incoming sounds.

Our new fossils show that the tympanic bone of pakicetids, unlike that of land mammals, was not connected rostro-medially to the periotic. This causes the thickened medial side of the tympanic (the pachyostotic involucrum) to form a loosely suspended centre of mass that could vibrate independently of the periotic. This would lead to enhanced transmission of bone-conducted sound when compared with generalized land mammals. An underwater sound transmission mechanism rooted in bone conduction is more sophisticated than that of a land mammal. A similar mechanism is probably used in modern phocid seals^{16,17}, in which the very heavy ossicles form the independently vibrating mass^{16,18} (Fig. 3, seal clusters), as well as in some subterranean mammals¹⁹.

New fossils for remingtonocetid and protocetid whales indicate that their sound transmission mechanism (Fig. 2c) combines aspects of those of pakicetids and modern odontocetes. In modern odontocetes, sound is transmitted to the middle ear by means of a large mandibular fat pad and received by the lateral wall of the tympanic bone (the tympanic plate), not the tympanic membrane^{20,21}. Sound passes through the odontocete middle ear by means of two lever-arm systems, and the freely suspended involucrum forms the axis of one of these^{10,12}.

Whereas pakicetids retain the land mammal morphology of the mandible, remingtonocetids and protocetids have a large mandibular foramen that indicates that the mandibular fat pad was present, and that the tympanic plate was functional during underwater hearing. The contact between periotic and tympanic bones in remingtonocetids and protocetids is more reduced than in pakicetids, and the involucrum is not attached to the periotic. The tympanic membrane (as evidenced by the morphology of the malleus and tympanic ring) is not the flat structure of generalized mammals, but has an elongated conical shape intermediate between land mammals and modern cetaceans. The size and shape of the ossicles, and the relation between ossicular mass and tympanic plate area (the underwater sound input area, Fig. 3) in these Eocene whales is similar to that in modern whales. Presence of the large mandibular foramen, suspension of the involucrum, and size and

shape of the ossicles indicate that the modern whale underwater sound transmission mechanism was realized.

Airborne sound reached the ear of remingtonocetids and protocetids through the external auditory meatus, a structure that is not patent in modern odontocetes. The remingtonocetid and protocetid malleus and incus are heavy and resemble those of modern cetaceans. The heavy ossicles imply that the frequency range for airborne sound was shifted to lower frequencies¹⁷. The malleus lacks a functional manubrium and is synostosed to the tympanic bone. These features indicate that reception of high-frequency airborne sound using the generalized mammalian system was present but poor.

In addition to the presence of the external auditory meatus, the remingtonocetid and protocetid ear differs from that of modern toothed whales in that it lacks air-filled sinuses that isolate the ear acoustically. The dorsal side of the ear (periotic) is dome-shaped (Fig. 2c) and fits in a dome-shaped depression in the skull (squamosal). These bones are not synostosed, but were probably connected by ligaments. These ligaments provide some, but very incomplete, acoustic insulation from bone-conducted sound. Hence, directional hearing was poor underwater. Transmission of airborne sound in remingtonocetids and protocetids was severely compromised by modifications of the ear for underwater hearing.

Basilosauroids follow protocetids chronologically, and are higher than protocetids on the cladogram²² (Fig. 4). The outer, middle and inner ear of basilosauroids is well-known^{4,13,23–25} and is functionally modern. A major functional innovation with respect to remingtonocetids and protocetids is that basilosauroid ears are acoustically isolated by the insertion of air-filled sinuses between the ear and the skull^{13,23}. Unlike modern cetaceans, basilosauroids have a large external auditory meatus that may have been patent. However, these whales were obligately marine²² and airborne sound was probably unimportant for them.

The history of Eocene cetaceans, as documented by our new fossils, shows that different organ systems followed a variety of evolutionary trajectories to reach modern morphologies and proportions. The dentition³ and the osmoregulatory system²⁶ show

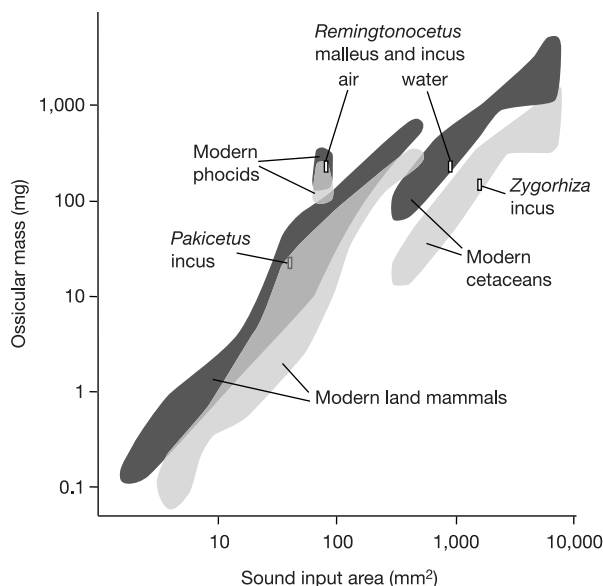


Figure 3 Bivariate plot of ossicular mass versus sound input area. Outline envelopes are for a representative sample of land mammals, phocid seals and modern cetaceans (see Methods and Supplementary Information). Ossicular mass is the combined mass of the malleus and the incus (dark grey) and the individual mass of the incus (light grey).

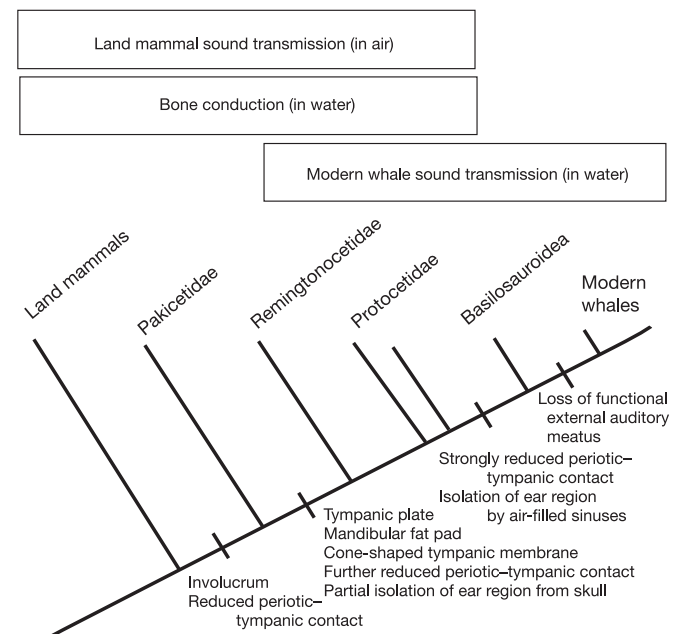


Figure 4 Generalized phylogeny for the cetaceans in this paper, with the characters relevant to hearing indicated at the nodes where they appeared. Boxes indicate the different sound transmission mechanisms used by these whales. The generalized mammalian sound transmission system was present in the cetacean sister group, here indicated as 'land mammals'.

intermediates. The organ of balance attained its modern proportions quickly and seemingly without transitional forms²⁷, and the locomotor system underwent a protracted and complex sequence of changes, representing a variety of functional patterns^{3,28}. The changes of the ear were complex because the physical properties of sound in air are very different from those in water. As a result, cetaceans totally reorganized sound transmission through the outer and middle ear (Fig. 4), while keeping the cochlea in the inner ear relatively unchanged.

This macroevolutionary change in the ear of cetaceans was nearly completed in four to seven million years. The functional end members of this evolutionary sequence—generalized mammalian hearing and modern whale hearing¹²—are widely distributed in modern forms; they can be seen as evolutionarily stable configurations²⁹. We propose that in these evolutionary end members all parts of the ear are optimized for collaboration with each other and are kept stable by internal selection²⁹. As the ancestors of cetaceans took to the water, the environmental tolerances of this system were exceeded and natural selection for the transmission of waterborne sound played its part. In pakicetids this resulted in a functional trade-off for the existing transmission mechanism, in which anatomical elements used for generalized sound transmission are now also important in bone-conducted hearing. The result is a sound transmission mechanism that works in air and in water, but performs poorly in both when compared with either land mammals or modern whales. A new evolutionarily stable configuration was not reached until new anatomical elements assumed a function in hearing in remingtonocetids and protocetids, and some old elements were eliminated as sound transmitters. The mandible is one of the most important elements added to the cetacean hearing mechanism and could be considered a keystone character²⁹ that catalysed the major transformation of the cetacean ear. □

Methods

In Fig. 3, sound input areas are those parts of the middle ear that are the sound receiver. For modern land mammals and phocid seals the sound input area is the tympanic membrane area, whereas for modern cetaceans it is the tympanic plate area. The tympanic membrane and tympanic plate are functionally analogous, but they are not evolutionarily homologous. For *Pakicetus* (H-GSP 91035), the tympanic membrane is the sound receiver; its area was plotted on the *x* axis. *Remingtonocetus* (RUSB 2914) used the tympanic membrane for receiving airborne sound, but the tympanic plate for receiving waterborne sound; hence, it was plotted twice in this figure. *Indocetus* is not shown because no incus is known for this taxon. The point for *Zygorhiza* was made on the basis of LSUMG V160A. The areas of the tympanic membrane and tympanic plate were determined using a method described elsewhere^{16,30}. Fossil taxa are plotted as ranges on the *y* axis because their density can only be estimated. The lower and upper range for the ossicular mass presented were calculated by multiplying the measured volumes by the minimum and maximum ossicular densities; that is, those for land mammals (2.0 g cm^{-3}) and modern cetaceans (2.7 g cm^{-3}), respectively³⁰. For data sources see Supplementary Information.

Received 13 February; accepted 7 June 2004; doi:10.1038/nature02720.

- Berta, A. What is a whale? *Science* **263**, 180–181 (1994).
- Fordyce, E. & de Muizon, C. in *Secondary Adaptation of Tetrapods to Life in Water* (eds Mazin, J. -M. & de Buffrénil, V.) 169–233 (Verlag Dr. Friedrich Pfeil, München, 2001).
- Thewissen, J. G. M. & Williams, E. M. The early radiations of Cetacea (Mammalia): Evolutionary pattern and developmental correlations. *Annu. Rev. Ecol. Syst.* **33**, 73–90 (2002).
- Fleischer, G. Evolutionary principles of the mammalian middle ear. *Adv. Anat. Embryol. Cell Biol.* **55**, 1–70 (1978).
- Lancaster, W. C. The middle ear of the Archaeoceti. *J. Vertebr. Paleontol.* **10**, 117–127 (1990).
- Oelschläger, H. A. in *Sensory Abilities of Cetaceans* (eds Thomas, J. A. & Kastlein, R. A.) 137–162 (Plenum, New York, 1990).
- Gingerich, P. D., Wells, N. A., Russell, D. E. & Shah, S. M. I. Origin of whales in epicontinental remnant seas: New evidence from the early Eocene of Pakistan. *Science* **220**, 403–406 (1983).
- Thewissen, J. G. M. & Hussain, S. T. Origin of underwater hearing in whales. *Nature* **361**, 444–445 (1993).
- Nummela, S., Reuter, T., Hemilä, S., Holmberg, P. & Paukku, P. The anatomy of the killer whale middle ear (*Orcinus orca*). *Hear. Res.* **133**, 61–70 (1999).
- Hemilä, S., Nummela, S. & Reuter, T. A model of the odontocete middle ear. *Hear. Res.* **133**, 82–97 (1999).
- Hemilä, S., Nummela, S. & Reuter, T. Modeling whale audiograms: effects of bone mass on high-frequency hearing. *Hear. Res.* **151**, 221–226 (2001).
- Møller, A. R. in *Handbook of Sensory Physiology* (eds Keidel, W. D. & Neff, W. D.) Vol. V/1, 491–517 (Springer, Berlin, 1974).
- Luo, Z. in *The Emergence of Whales* (ed. Thewissen, J. G. M.) 269–301 (Kluwer Academic/Plenum, New York, 1998).

- Lombard, R. E. & Hetherington, T. E. in *The Skull* (eds Hanken, J. & Hall, B. K.) 241–302 (Univ. Chicago Press, Chicago, 1993).
- Tonndorf, J. A new concept of bone conduction. *Arch. Otolaryng.* **87**, 49–54 (1968).
- Nummela, S. Scaling of the mammalian middle ear. *Hear. Res.* **85**, 18–30 (1995).
- Hemilä, S., Nummela, S. & Reuter, T. What middle ear parameters tell about impedance matching and high-frequency hearing. *Hear. Res.* **85**, 31–44 (1995).
- Wyss, A. R. The walrus auditory region and the monophyly of pinnipeds. *Am. Mus. Novit.* **2871**, 1–31 (1987).
- Mason, M. J. Middle ear structures in fossorial mammals: a comparison with non-fossorial species. *J. Zool.* **255**, 467–486 (2001).
- Norris, K. S. in *Marine Bio-Acoustics* (ed. Tavolga, W. N.) 317–336 (Pergamon, Oxford, 1964).
- Brill, R. L., Moore, P. W. B. & Dankiewicz, L. A. Assessment of dolphin (*Tursiops truncatus*) auditory sensitivity and hearing loss using headphones. *J. Acoust. Soc. Am.* **109**, 1717–1722 (2001).
- Uhen, M. D. in *The Emergence of Whales* (ed. Thewissen, J. G. M.) 29–61 (Kluwer Academic/Plenum, New York, 1998).
- Luo, Z. & Gingerich, P. D. Terrestrial Mesonychia to aquatic Cetacea: Transformation of the basicranium and evolution of hearing in whales. *Univ. Mich. Pap. Paleontol.* **31**, 1–98 (1999).
- Fleischer, G. Hearing in extinct cetaceans as determined by cochlear structure. *J. Paleontol.* **50**, 133–152 (1976).
- Ketten, D. R. in *The Evolutionary Biology of Hearing* (eds Webster, D. B., Fay, R. R. & Popper, A. N.) 717–750 (Springer, New York, 1992).
- Thewissen, J. G. M. *et al.* Evolution of cetacean osmoregulation. *Nature* **381**, 379–380 (1996).
- Spoor, F., Bajpai, S., Hussain, S. T., Kumar, K. & Thewissen, J. G. M. Vestibular evidence for the evolution of aquatic behaviour in early cetaceans. *Nature* **417**, 163–166 (2002).
- Gingerich, P. D. Land-to-sea transition in early whales: evolution of Eocene Archaeoceti (Cetacea) in relation to skeletal proportions and locomotion of living semiaquatic mammals. *Paleobiology* **29**, 429–454 (2003).
- Wagner, G. P. & Schwenk, K. Evolutionarily stable configurations: functional integration and the evolution of phenotypic stability. *Evol. Biol.* **31**, 155–217 (2000).
- Nummela, S., Wagar, T., Hemilä, S. & Reuter, T. Scaling of the cetacean middle ear. *Hear. Res.* **133**, 71–81 (1999).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank E. Blum, K. Grecco and F. Spoor for help with CT scans; W. Lancaster for information about basilosauroid ossicles; M. Filon for artwork; and S. Hemilä, S. Madar, T. Reuter and L. Sundström for commenting on the manuscript. J. Schiebout (Louisiana State University) and H. Baagøe and M. Andersen (Zoological Museum, University of Copenhagen) loaned specimens. The Geological Survey of Pakistan assisted in the collection and loan of some of the fossils. Financial support was provided to J.G.M.T. by the National Science Foundation and to S.B. by the Department of Science of Technology of India.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.G.M.T. (thewis@neoucom.edu).

Pinyon jays use transitive inference to predict social dominance

Guillermo Paz-y-Miño C¹, Alan B. Bond¹, Alan C. Kamil^{1,2} & Russell P. Balda³

¹Center for Avian Cognition, School of Biological Sciences and ²Department of Psychology, University of Nebraska, Lincoln, Nebraska 68588, USA

³Department of Biological Sciences, Northern Arizona University, Flagstaff, Arizona 86011, USA

Living in large, stable social groups is often considered to favour the evolution of enhanced cognitive abilities, such as recognizing group members, tracking their social status and inferring relationships among them^{1–4}. An individual's place in the social order can be learned through direct interactions with others, but conflicts can be time-consuming and even injurious. Because the number of possible pairwise interactions increases rapidly with group size, members of large social groups will benefit if they can make judgments about relationships on the basis of indirect evidence⁵. Transitive reasoning should therefore be particularly important for social individuals, allowing assessment of relationships from observations of interactions among others. Although

a variety of studies have suggested that transitive inference may be used in social settings^{6–10}, the phenomenon has not been demonstrated under controlled conditions in animals. Here we show that highly social pinyon jays (*Gymnorhinus cyanocephalus*) draw sophisticated inferences about their own dominance status relative to that of strangers that they have observed interacting with known individuals. These results directly demonstrate that animals use transitive inference in social settings and imply that such cognitive capabilities are widespread among social species.

Pinyon jays are among the most social of North American corvids. They live in large permanent flocks of up to 500 individuals, breed colonially and establish long-term multigenerational relationships with linear dominance hierarchies^{11,12}. In operant experiments using coloured stimuli, pinyon jays track implicitly ordered dyadic relationships more accurately and display more robust transitive inferences than western scrub jays (*Aphelocoma californica*), a closely related, less social species¹³. This suggests that the differing demands of the social systems in the two species may have selected for differential cognitive abilities¹⁴, but there has been no direct evidence that pinyon jays or any other non-human species use transitive inference to make social judgments.

Sixteen adult male pinyon jays, sexed by DNA analysis, were captured in northern Arizona, housed individually and kept mildly hungry (at 90% of their free-feeding weight) by controlled daily feedings. They were divided into three groups such that, although some birds may have known each other from the wild, no birds from separate groups had been in direct contact for at least five years (see Methods). Within the three groups dominance relationships were established in a series of six 5-min staged encounters between each of the 36 possible pairs of individuals (Fig. 1a). All sessions were recorded on digital videotape and the frequencies of five behavioural events—one dominant display (stare at) and four submissive ones (look away, crouch, chin-up and beg)^{11,12}—were determined for each bird.

From each encounter session, we extracted two weighted indices for each bird (one of dominant and one of subordinate behaviours) on the basis of the event frequencies. Weights were obtained by canonical discriminant analysis on a subset of dyads in which the relationships were clear and unequivocal (see Methods). The difference between dominant and subordinate behaviours for a given individual provided a direct measure of the strength of its relative social status (see Methods). Using the average differences in relative social status between individuals in all 36 dyads over the last three encounters, we constructed inferred within-group dominance hierarchies and coded the bird designations accordingly. In all three groups the hierarchies were linear and fully transitive (group 1, $A > B > C > D > E > F$; group 2, $1 > 2 > 3 > 4 > 5 > 6$; and group 3, $P > Q > R > S$). For 62% of the dyads, relative social status was consistent from the first or second encounter. In those dyads in which both birds were low-ranking, however, the process often took much longer: over 25% of the dyads showed transient status reversals as late as the fourth or fifth encounter.

Once within-group dominance relationships had been established, we conducted a limited set of cross-group dominance encounters between similarly ranked individuals to establish a basis for predicting the relationships between other, untested cross-group dyads. During the cross-group encounters we avoided using birds at the top or bottom of their group hierarchies, because the next stage of the experiment ('exhibition' encounters, below) required birds that could both win and lose encounters with members of their group. There were 32 possible cross-group dominance relationships, of which we determined eight (see Methods). The information from these cross-group pairings plus the knowledge of each within-group hierarchy made it possible to establish our experimental and control conditions in the next stage of the experiment.

We designed a set of 12 instances that tested the ability of pinyon

jays to draw social inferences. In each instance, an observer bird watched a series of 'exhibition' encounters between another bird (the 'demonstrator') and two different opponents (Fig. 1b, c). Although the observer had never previously interacted with the demonstrator, the results of the cross-group dominance tests predicted that the demonstrator should dominate the observer. On each of three consecutive days, the observer watched his demonstrator lose encounters with one opponent and win encounters with another, giving a total of six exhibition encounters per observer/demonstrator dyad. In all cases, the outcomes of the exhibition

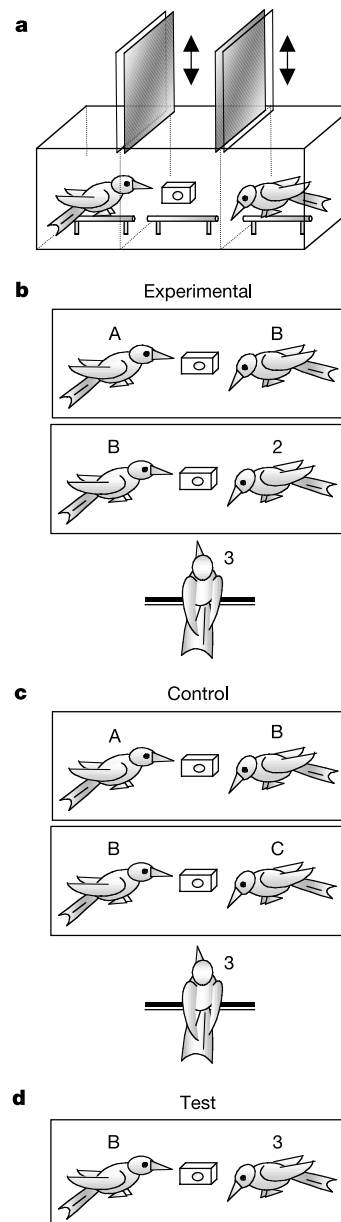


Figure 1 Dominance encounters were conducted in a transparent acrylic chamber (100 × 40 × 40 cm) separated into three compartments by sliding opaque and transparent dividers. On each trial, a feeder box in the central compartment was baited with a single unshelled peanut. **a**, To establish relative dominance, all dyads within each of the social groups were tested. Birds were placed in each of the two end chambers and were released simultaneously to compete over access to the peanut. **b, c**, Next, exhibition encounters were observed by a third bird whose prior knowledge of contestants was systematically controlled. If bird 3 was in the experimental condition, for example, he watched $A > B$, $B > 2$ (only 2 known to 3). If bird 3 was in the control condition, he watched $A > B$, $B > C$ (all strangers to 3). **d**, In either case, following these exhibition encounters 3 was tested with B.

encounters coincided with our predictions. This sequentially balanced design was employed to ensure the use of transitive inference on the basis of observations of encounters between specific individuals, rather than simply the general effects of having seen another animal win or lose^{15–17}.

In the exhibition encounters, the observer's prior knowledge about one of the demonstrator's opponents was varied systematically. In experimental instances the demonstrator and one opponent were strangers, but the opponent that repeatedly lost to the demonstrator was familiar to the observer, because that opponent had dominated the observer in earlier staged encounters. For example, suppose that cross-group testing had established that B was dominant to 2. We could then use 3 in an experimental treatment by allowing him to observe that $A > B$ and $B > 2$ (Fig. 1b). In this case, A and B would be strangers to 3, but the relationship $2 > 3$ would have been established in earlier within-group interactions. Using transitive reasoning, 3 should expect to be subordinate to B. In contrast, in control instances the demonstrator and both of his opponents were all strangers to the observer. If 3 were assigned to the control condition, he would watch $A > B$ and $B > C$, all unknown to 3 (Fig. 1c). Therefore, 3 could not subsequently use transitive reasoning to predict his status relative to B.

Following the exhibition encounters, each observer was given six 5-min staged encounters with his demonstrator (testing, in our example, B versus 3; Fig. 1d), and the behaviour of both participants in these test trials was evaluated for evidence of transitive social inference. We conducted six sets of experimental and six sets of control exhibition and test encounters, each using different observer/demonstrator dyads (see Methods). Because the data violated assumptions of parametric analysis, differences between groups were analysed using Wilcoxon two-sample exact probability tests¹⁸. Dyads were assigned to treatments in a balanced fashion,

such that experimental and control dyads had statistically indistinguishable differences in relative social status scores ($W^+ = 44$, $P = 0.48$).

If experimental observers use transitive social inference to predict their relationship to the demonstrator, they should exhibit higher initial levels of submissive behaviour than control observers. This proved to be the case. During the first minute of the first encounter experimental observers displayed subordinate levels that were nearly four times as high as those of controls (Fig. 2a). The effect faded rapidly, however, and there were no significant treatment differences in subordinate behaviour between observers in subsequent intervals or in later encounters. Experimental observers also displayed lower initial levels of dominant behaviours than controls, though the effect was only statistically significant in the second minute of the first encounter (Fig. 2c). There were no significant differences in subordinate behaviour between demonstrators paired with experimental or control observers (Fig. 2b), but experimental demonstrators showed higher levels of dominance than controls (Fig. 2d) during the second and third minutes of the first encounter, apparently in response to the initially higher level of subordinate behaviour displayed by experimental observers. As with the observer's behaviour, the effect on the demonstrator's dominance was transient and no significant differences were apparent in later encounters.

In addition to win or lose information, exhibition encounters also provided information about the relative disparity between a demonstrator and his opponents. Observers could, therefore, have made more subtle, graded assessments of their relative status, inferring the probable degree of difference in dominance between themselves and the demonstrator. To determine whether observers made use of this information, we calculated the difference in relative dominance for each demonstrator between the exhibition encounters he won and those he lost. We tested this measure as a predictor of observer subordinate behaviour during the first test encounter. Whereas there was a strong positive relationship for experimental observers, there was no such relationship for control birds (Fig. 3). These findings suggest that observers estimated the actual disparity in their dominance status relative to the demonstrator, but only when they knew the losing opponent.

The results are fully in accord with the hypothesis that pinyon jays use transitive reasoning to make inferences of relative dominance.

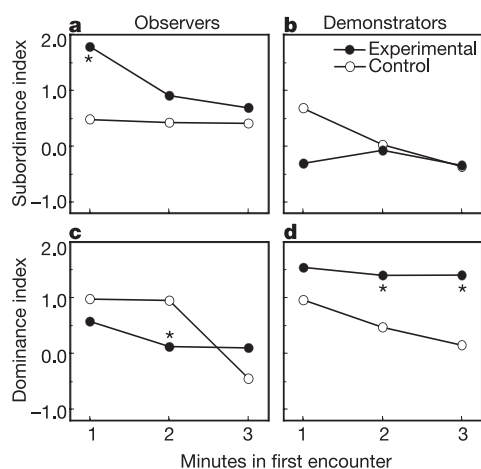


Figure 2 Mean dominance and subordinate indices for each of the first 3 min of the first test session. **a–d**, Behaviour of the observers (**a**, **c**) and the corresponding demonstrators (**b**, **d**); filled circles correspond to experimental birds, open circles to control birds. **a**, Experimental observers were significantly more subordinate during the first minute than control observers ($W^+ = 53$, $P = 0.026$), but this effect disappeared after the first minute ($W^+ = 41$, $P = 0.41$). **b**, There were no significant differences in subordinate behaviour between demonstrators paired with experimental or control observers ($W^+ \leq 47$, $P \geq 0.12$). **c**, Experimental observers displayed lower levels of dominant behaviours than controls, though the difference was only significant during the second minute of the first encounter (second: $W^+ = 53$, $P = 0.022$; first and third: $W^+ \leq 47$, $P \geq 0.12$). Because of their higher level of subordinate behaviour in the first minute (**a**), relative social status (= dominance – subordinate) was markedly lower for experimental observers during the first minute of the first encounter ($W^+ = 54$, $P = 0.015$). **d**, Demonstrators showed higher levels of dominance than controls, but the effect was significant only in the second and third minutes of the first encounter (first: $W^+ = 42$, $P = 0.34$; second and third: $W^+ \geq 52$, $P \leq 0.039$).

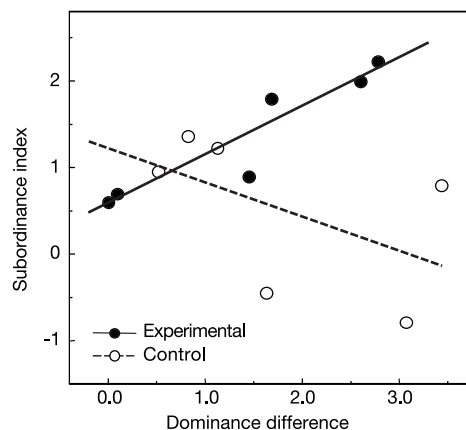


Figure 3 Subordinate index for observers during the first 5-min test encounter with the demonstrator. The subordinate index is a function of the mean difference in dominance between demonstrators and opponents during the exhibition encounters. Experimental dyads are shown as filled circles with a continuous regression line; control dyads as open circles with a dashed regression line. There was a strong positive relationship ($F_{1,5} = 26.64$, $P = 0.007$, $r^2 = 0.87$) for experimental birds, but none for control birds ($F_{1,5} = 1.55$, $P > 0.25$, $r^2 = 0.28$), and analysis of covariance (ANCOVA) indicated that the regression slopes for the two treatments were significantly different ($F_{1,2} = 8.01$, $P = 0.022$).

Jays that had previously interacted with one of the birds they observed drew inferences about their rank relative to the demonstrator, and showed a graded, quantitative response based on their observations. Jays that observed very similar interactions, but had never interacted directly with any of the birds they observed, failed to show either effect. This pattern rules out alternative general explanations, such as badges of status¹⁹ or dispositional responses to seeing another bird win or lose^{15–17}. This work constitutes a direct demonstration of transitive inference in social settings, and supports the hypothesis that social complexity provided a crucial context for the evolution of cognitive abilities. □

Methods

Test procedures

To familiarize the birds with the apparatus, each jay was placed alone in one of the end compartments of the encounter chamber. After 30 s the dividers were raised and the bird was allowed to explore the apparatus until it discovered and consumed a peanut. Each bird received six such familiarization trials before beginning staged encounters. During staged encounters (Fig. 1a), each member of a dyad was initially placed in one of the end chambers (randomly selected). After 10 s the opaque divider was lifted, providing visual contact between dyad members through the second, transparent divider. After an additional 10 s the transparent divider was lifted, giving the birds simultaneous access to the central contest area. To facilitate recognition of individuals for video scoring, one of the dyad members in each encounter was marked on the wing primaries with water-soluble white paint. After the encounter the paint was removed.

Group formation and selection of pairs for testing

Our experimental design required sets of birds of relatively similar rank who were unknown to each other, but whose relative dominance could be predicted accurately. We first divided the birds into three groups; two groups of six birds and one group of four. The small size of the groups minimized the possibility of nonlinear relationships. Once the within-group hierarchies were established, we then determined eight of the 32 possible cross-group dominance relationships by pairing the second-, third- and fourth-ranked birds in group 1 with those of the same rank in group 2, and the second- and third-ranked birds in group 3 with the second- and third-ranked birds in both groups 1 and 2. The outcomes of these within- and cross-groups dominance encounters were subsequently used to select sets of observers and demonstrators.

During exhibition sessions the demonstrator was paired with two other birds, one dominant and one subordinate to the demonstrator. In experimental conditions one of these other birds had to be a stranger to the observer and the other a known dominant. In control conditions both birds had to be strangers to the observer. In addition, because the experimental design required birds that could both win and lose encounters with members of their group, birds at the top or bottom of their group hierarchies were not used as observers. These constraints limited the number of possible pairings that could be generated, with the result that some individuals were used in more than one trial. Nine observers and six demonstrators participated in the six experimental and six control pairings.

To control for prior experience in winning and losing, we arranged daily maintenance encounters between each of the demonstrators and observers and members of their own groups. During the three weeks before testing, each observer had an average of 15 encounters with five other birds, of which he won 47%; for the four observers that were tested more than once, at least two months passed between successive trials. In the same time period, each demonstrator had an average of 16 prior encounters with four other birds, of which he won 56%; for the three demonstrators tested more than once, at least 10 days passed between successive trials.

Behavioural indices

Because display behaviour is often a more reliable indicator of dominance than gaining access to food²⁰, we used relative frequencies of behavioural acts to assess dominance. To obtain an empirically valid index of relative dominance in which the contributions of the different behavioural events were appropriately weighted, we first calculated (for each individual in each encounter) the difference between the raw counts of dominant and subordinate actions divided by their sum. Differences between dyad members in the value of this ratio, which weighted all action patterns equally, provided an initial approximate measure of relative dominance. From the 36 within-group dyads in the study, we extracted a set of 15 exemplars, dyads in which the mean of this ratio (averaged over all six encounters) was larger than 0.5 and in which one of the dyad members consistently dominated in all six encounters. Because some behaviours are better indicators of social status than others, however, a simple sum of event frequencies is often misleading as an indicator of relative dominance. To obtain a more sensitive measure, we subjected the raw counts from the last three encounters from each exemplar to canonical discriminant analysis¹⁸, which produces the weighted linear combination of standardized variables that best distinguishes between data classes. In the final configuration, three variables—the frequencies of stare at and look away, and the sum of the frequencies of the three other submissive displays—were log-transformed, standardized and combined into two weighted discriminant functions that constituted dominance and subordination indices. The difference between dominance and subordination provided a direct measure of each individual's relative social status (relative social status = dominance – subordination), and in this combination the discriminant functions correctly categorized 93% of the encounters in the exemplar data set.

Received 7 April; accepted 3 June 2004; doi:10.1038/nature02723.

1. Jolly, A. Lemur social behavior and primate intelligence. *Science* **153**, 501–506 (1966).
2. Humphrey, N. K. in *Growing Points in Ethology* (eds Bateson, P. & Hinde, R. A.) 303–317 (Cambridge Univ. Press, Cambridge, 1976).
3. Kummer, H., Daston, L., Gigerenzer, G. & Silk, J. in *Human by Nature: Between Biology and the Social Sciences* (ed. Weingart, P. et al.) 157–179 (L. Erlbaum, Hillsdale, 1997).
4. de Waal, F. B. M. & Tyack, P. L. *Animal Social Complexity* (Harvard Univ. Press, Cambridge, 2003).
5. Seyfarth, R. M. & Cheney, D. L. in *Animal Social Complexity* (eds de Waal, F. B. M. & Tyack, P. L.) 207–229 (Harvard Univ. Press, Cambridge, 2003).
6. Cheney, D. L. & Seyfarth, R. M. Vocal recognition in free-ranging vervet monkeys. *Anim. Behav.* **28**, 362–367 (1980).
7. Holekamp, K. E. et al. Vocal recognition in the spotted hyena and its possible implications regarding the evolution of intelligence. *Anim. Behav.* **58**, 383–395 (1999).
8. Silk, J. B., Seyfarth, R. M. & Cheney, D. L. The structure of social relationships among female baboons. *Behaviour* **136**, 679–703 (1999).
9. Peake, T. M., Terry, A. M. R., McGregor, P. K. & Dabelsteen, T. Do great tits assess rivals by combining direct experience with information gathered by eavesdropping? *Proc. R. Soc. Lond. B* **269**, 1925–1929 (2002).
10. Bergman, T. J., Beehner, J. C., Cheney, D. L. & Seyfarth, R. M. Hierarchical classification by rank and kinship in baboons. *Science* **302**, 1234–1236 (2003).
11. Balda, R. P. in *The Birds of North America* No. 605 (eds Poole, A. & Gill, F.) 1–32 (Birds of North America Inc., Philadelphia, 2002).
12. Marzluff, J. M. & Balda, R. P. *The Pinyon Jay* (T. & A.D. Poyser, London, 1992).
13. Bond, A. B., Kamil, A. C. & Balda, R. P. Social complexity and transitive inference in corvids. *Anim. Behav.* **65**, 479–487 (2003).
14. Balda, R. P., Kamil, A. C. & Bednekoff, P. A. in *Current Ornithology* **13** (eds Nolan, V. & Ketterson, E. D.) 33–66 (Plenum, New York, 1996).
15. Chase, I., Bartolomeo, C. & Dugatkin, L. A. Aggressive interactions and inter-contest interval: how long do winners keep winning? *Anim. Behav.* **48**, 393–400 (1994).
16. Hogue, M.-E., Beaugrand, J. P. & Laguë, P. C. Coherent use of information by hens observing their former dominant defeating or being defeated by a stranger. *Behav. Processes* **38**, 241–252 (1996).
17. Oliveira, R. E., McGregor, P. K. & Latruffe, C. Know thine enemy: fighting fish gather information from observing conspecific interactions. *Proc. R. Soc. Lond. B* **265**, 1045–1049 (1998).
18. SAS v.8 (SAS Institute Inc., Cary, North Carolina, 2000).
19. Rohwer, S. The evolution of reliable and unreliable badges of fighting ability. *Am. Zool.* **22**, 531–546 (1982).
20. Huntingford, F. A. & Turner, A. K. *Animal Conflict* (Chapman and Hall, London, 1987).

Acknowledgements We thank N. Howe and E. A. Simpson for assistance in data collection and B. Luke Stafford for help in designing Fig. 1. Supported by University of Nebraska Research Enhancement Funds and an NSF grant to Northern Arizona University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.B.B. (abond@unl.edu) or A.C.K. (akamil@unl.edu).

Complex auditory behaviour emerges from simple reactive steering

Berthold Hedwig & James F. A. Poulet

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

The recognition and localization of sound signals is fundamental to acoustic communication^{1,2}. Complex neural mechanisms are thought to underlie the processing of species-specific sound patterns even in animals with simple auditory pathways^{3,4}. In female crickets, which orient towards the male's calling song, current models propose pattern recognition mechanisms based on the temporal structure of the song^{5–7}. Furthermore, it is thought that localization is achieved by comparing the output of the left and right recognition networks, which then directs the female to the pattern that most closely resembles the species-specific song^{8–10}. Here we show, using a highly sensitive method for measuring the movements of female crickets, that when walking and flying each sound pulse of the communication signal releases a rapid steering response. Thus auditory orientation emerges from reactive motor responses to individual sound

pulses. Although the reactive motor responses are not based on the song structure, a pattern recognition process may modulate the gain of the responses on a longer timescale. These findings are relevant to concepts of insect auditory behaviour and to the development of biologically inspired robots performing cricket-like auditory orientation^{11–13}.

Many animals, including frogs and insects, communicate with stereotyped patterns of constant frequency sound pulses¹⁴, making them suitable models to investigate the neuronal mechanisms underlying acoustic pattern recognition and localization^{15,16}. Male crickets (*Gryllus bimaculatus*) produce long-lasting calling songs, made up of chirps of four to five sound pulses^{17,18} that are repeated at 2–3 Hz. Female crickets are attracted by this calling song and walk^{4,19} or fly^{20,21} towards singing males. Auditory processing takes place in a simple pathway^{15,22}. When exposed to artificial songs, females perform phonotactic walking on trackballs, which allows quantitative analysis of their behaviour^{23–25}. With such experiments, pattern recognition is inferred from the animal's auditory orientation⁴. Previous investigators relied on high inertia trackballs and monitored the animal's walking speed averaged over 500–1,000 ms. Here we used a new, highly sensitive trackball (Fig. 1a), which measured the path of tethered stationary walking crickets with a temporal resolution of 0.3 ms (see Methods). Modulations of walking speed by individual steps could be recorded (Fig. 1b) and it revealed so far undetected rapid steering movements.

Auditory pattern recognition in female crickets is tuned to the species-specific song structure and may be achieved by template matching⁵, temporal band-pass filtering⁶ or cross-correlation analysis⁷. To gain insight into auditory orientation, we analysed the steering behaviour to split-song patterns²⁶. A chirp with six sound pulses (duration 21 ms, interval 21 ms) was split so that every two consecutive pulses were presented from opposite sides (Fig. 2a, b; top) at $\pm 45^\circ$ from the animal's longitudinal axis. At the onset of sound, the animals started oriented walking with a speed of $5\text{--}7\text{ cm s}^{-1}$. As a measure of acoustic orientation we calculated the cricket's overall lateral deviation towards the left or right side from a forward path (Fig. 2a; middle), and as a measure of steering we calculated the actual lateral steering velocity by which the animals steered to the left and right side (Fig. 2a; bottom). All females tested ($N = 10$) deviated towards the speaker presenting the four sound pulses. On the basis of previous models of pattern recognition we might conclude that the animals oriented towards the 'better' sound pattern, comprising four pulses^{9,10}. The concomitant lateral steering velocity, however, oscillated around zero and was directed both to the left and to the right side (Fig. 2a; lower trace). Conspicuous peaks in the lateral steering velocity were closely linked to the occurrence of sound pulses (Fig. 2b) so that each pair of sound pulses elicited a rapid steering transient with maximum peak velocities of 5 cm s^{-1} towards the side of the active speaker (Fig. 2b; lower trace). Plotting the lateral deviation of the animal at high resolution showed that each sound-evoked velocity-transient led to a corresponding deviation of the animal's path by about 1 mm towards the left or right side respectively. Therefore, overall the animal oriented towards the speaker presenting four pulses (Fig. 2b; middle). We quantified the steering behaviour and averaged the lateral deviation velocity signal over 400 chirps (Fig. 2c). Sound pulses presented from the left elicited a transient velocity peak directed to the left, and pairs of sound pulses from the right caused a transient lateral velocity component to the right. These steering responses occurred with a latency of only 55–60 ms and started during the second sound pulse of a pair.

To determine whether walking or flying crickets will also steer towards individual sound pulses we took the split-song model to its extreme and presented every other sound pulse from opposite directions. Averaging the lateral steering velocity of walking crickets ($N = 15$) clearly showed that the crickets rapidly turned towards the sound pulses presented from the left and right side in alternation

(Fig. 2d). Auditory steering responses occurred again after 55–60 ms and lasted for 42 ms, the duration of a pulse period. These reflex-like steering responses had very similar amplitudes and because they were directed in opposite directions they cancelled out and, as in crickets walking under open-loop conditions²⁶, resulted in a net walking direction midway between both speakers. Crickets not only walk but also fly towards singing males, steering with lateral movements of their abdomen²¹. We therefore tested the flight steering of the same crickets that had walked on the trackball. Females ($N = 6$) were exposed to a constant wind stream to elicit flying and their abdominal steering movements were measured with an optoelectronic camera²⁷ (see Methods). Upon acoustic stimulation the females produced rapid abdominal movements towards the side presenting the sound pattern²⁸. Averaging the responses towards the split-song model revealed that they followed the same time course as the steering responses during phonotactic walking (Fig. 2e). These results imply that cricket auditory orientation during walking and flight is based on rapid reactive steering towards individual pulses of the male's calling song.

On the basis of this finding we expected a clear relationship between the number of sound pulses perceived from each side and the overall lateral deviation in walking crickets. We therefore systematically varied the ratio of sound pulses presented from the left or the right. Chirps with six pulses were used and one, two or

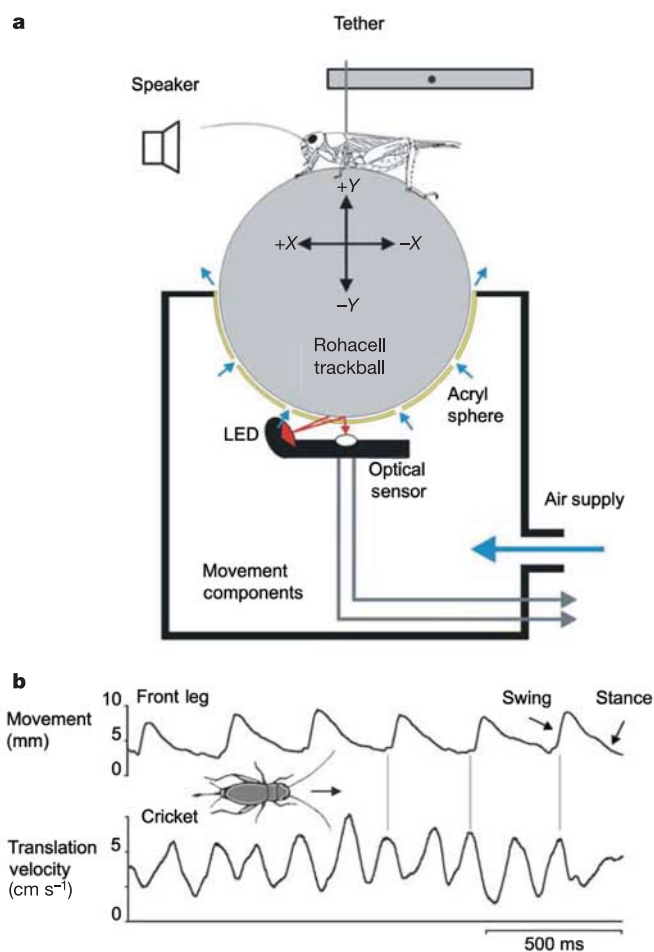


Figure 1 Trackball system for testing auditory orientation of crickets. **a**, A cricket is tethered on a trackball that is floating in an air stream. An optical sensor picks up the movements of the trackball, which is rotated by the cricket walking. **b**, Simultaneous recording of the stance–swing movements of one front leg femur (top, see Methods) and the translation velocity of the cricket (bottom). Owing to insects' tripod gait the translation velocity oscillated with twice the frequency of the front leg movements. Every second velocity peak is in phase with the swing phase of the front leg.

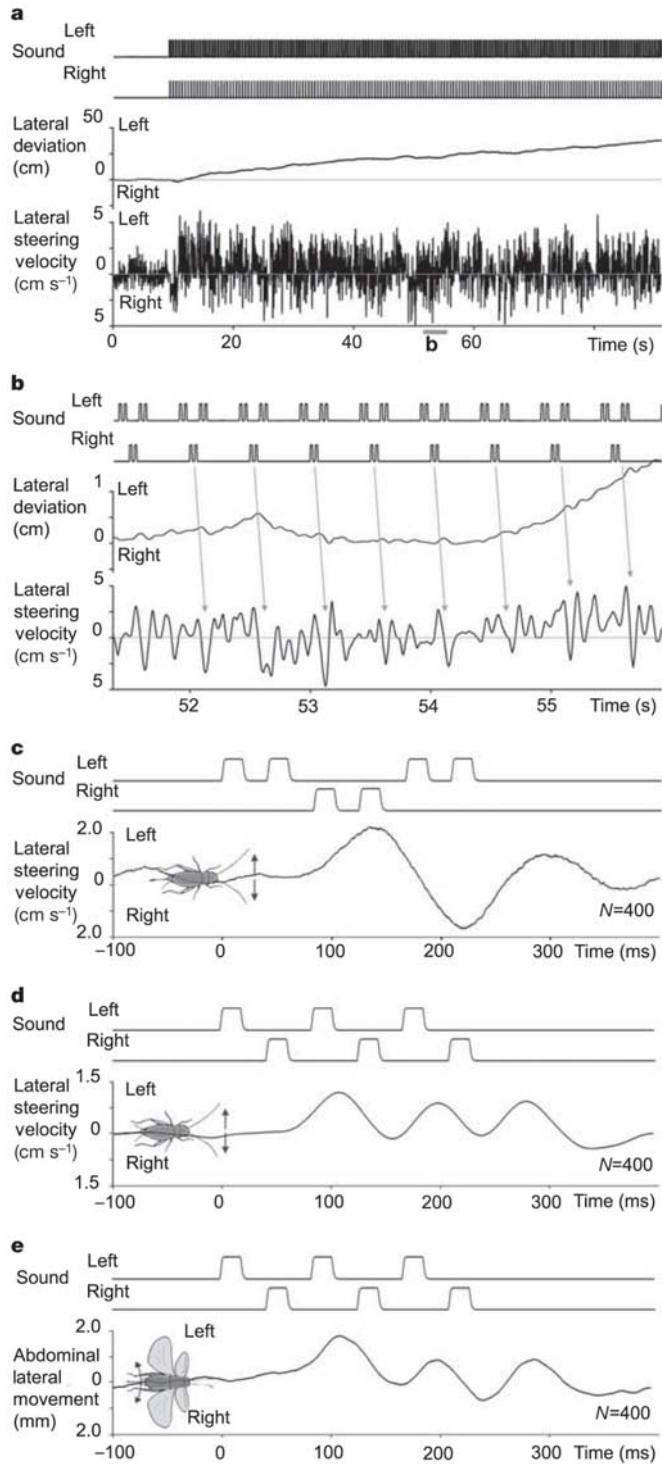


Figure 2 Cricket orientation to songs that are split between two speakers. **a**, The cricket walked towards the left speaker, which presented double the number of pulses. The lateral steering velocity oscillated around zero revealing steering transients to the left and right side. **b**, A high-resolution plot of the section labelled 'b' in **a** shows that each pair of sound pulses elicited a steering transient towards the corresponding speaker. **c**, Dynamics of rapid steering revealed by averaging the lateral steering velocity for 400 chirps. Each pair of sound pulses elicited a steering transient towards the side presenting the sound. **d**, **e**, Rapid steering in walking and flying crickets. Consecutive sound pulses were presented from alternating sides. The lateral steering velocity of the walking cricket (**d**) and the abdominal steering response of the same cricket during flight (**e**) demonstrate steering towards the sound pulses with the same temporal dynamic during walking and flight.

three pulses were randomly presented from the opposite direction (Fig. 3a; top). When all six sound pulses were presented from one side the animals oriented towards that speaker and the lateral steering velocity components were directed to that side (Fig. 3a; middle). As the number of sound pulses presented from the opposite side was increased, the animals deviated less to the side presenting the larger number of sound pulses. When both speakers presented an equal number of pulses the crickets walked forward and the lateral deviation was close to zero. Although the lateral deviation decreased, the females continued walking and covered an overall distance of 200–230 cm in 30 s in all tests (Fig. 3b). We quantified the overall lateral deviation as a function of the ratio of sound pulses presented (Fig. 3c). All females ($N = 11$) oriented towards the side presenting the larger number of sound pulses. With all sound pulses coming from one side they reached a control value of 100%. Because individual sound pulses elicited similar steering responses (Fig. 2d) we were able to calculate the expected performance of the animals and compare it with what we observed. At a ratio of 5:1 the females steered towards the speaker presenting five pulses with 68.9% (± 2.5 s.e.m.) of the reference value (expected 66.6%), at a ratio of 4:2 the deviation was 38.8% (± 3.2 s.e.m.) of the reference value (expected 33.3%) and at a ratio of 3:3 it was 6.0% (± 2.4 s.e.m.) with zero deviation expected. When exposed to two sound sources the animal's course depends on the ratio of pulses perceived from both sides and emerges as a result of numerous consecutive steering events towards individual sound pulses.

Consequently, we expected crickets to walk straight ahead when exposed to split-song patterns that contain identical numbers of sound pulses but when presented alone have a different attractive-

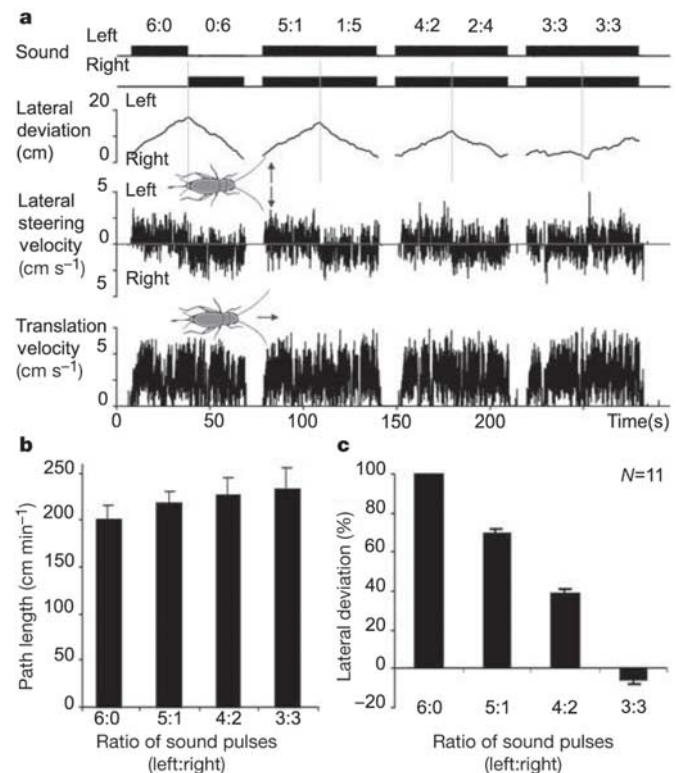


Figure 3 Cricket orientation to randomly split songs with different numbers of pulses presented from each side. **a**, The cricket deviated towards the side with more sound pulses per chirp; however, the animal produced steering transients directed to the left and right side. **b**, Path length walked, according to the ratio of sound pulses presented from opposite directions. **c**, The relative lateral deviation to the side presenting the larger number of sound pulses depends on the ratio of presented sound pulses. Error bars indicate s.e.m.

ness. A sound pattern of three pulses spaced with an interval of 147 ms is only weakly attractive and reached 14.0% (± 4.6) of the reference value (Fig. 4a, b; left). When the interval between the pulses was 21 ms, crickets clearly oriented with the reference response of 100% (Fig. 4a, b; middle). Steering towards the 21-ms-interval pattern, however, was reduced to 23.7% (± 4.8) when both patterns were presented simultaneously (Fig. 4a, b; right). An average of the lateral steering velocity shows that the animals consecutively steered towards the sound pulses of both patterns (responses to the left indicated by asterisks in Fig. 4c) with the initial response towards the first pulse of the 147-ms-pattern (Fig. 4c). Thus, orientation was not determined by the "better" pattern¹⁰, but by the number of sound pulses presented to each side. The result also indicates that the steering response towards the non-attractive pattern may have been altered when it was combined with the attractive pattern.

The phonotactic behaviour of female *G. bimaculatus* depends on the temporal structure of the sound pattern^{4,17,18}. However, when exposed to songs with the species-specific pulse duration and pulse interval, females rapidly steer towards individual sound pulses. This steering behaviour does not support current models for the interaction of pattern recognition and localization. It was proposed that the animals compare the outputs of two separate pattern recognizers on either side of the brain to calculate a steering direction for phonotactic walking^{8,9} and concluded that cricket auditory localization is governed by the rule 'turn to the better pattern'¹⁰. Our results contradict this model because orientation emerges from rapid steering towards individual sound pulses and has already been initiated before the central nervous system has had time to process the second pulse of a chirp. For the same reason, our results rule out the possibility that pattern recognition by band-pass filtering brain

neurons⁶, template matching⁵ or cross-correlation analysis⁷ is directly involved in rapid steering. Any of these processes based on the integration of the temporal pattern would require at least two consecutive sound pulses to determine the temporal structure of the perceived song and thus would be too slow to evoke the observed rapid steering responses. This does not exclude the possibility that the recognition process modulates the rapid steering responses over a longer time course.

Cricket auditory orientation emerges from reflex-like reactive steering towards individual sound pulses, and the path of the animals depends on the ratio of pulses perceived from the left and right side. The similarity of the steering responses in walking and flight indicates that the animals use the same strategy and pathway during both behaviours. Using reactive steering towards individual sound pulses, the animals elegantly solve a complex task of auditory orientation with a simple mechanism. This suggests a new concept for cricket phonotactic localization behaviour, based on reactive steering responses to sound pulses. Neuronal activity representing the sound pulses may directly act on the motor control networks²⁹. Low level sensory processing with a direct coupling of the sensory input to the motor output has been used in robots and produced cricket-like auditory behaviour^{11,12}. Our demonstration of reactive steering¹³ in cricket behaviour provides crucial information for the design of neuronal networks driving biologically inspired robots. □

Methods

Crickets

Female crickets (*G. bimaculatus* de Geer) were taken as last instars from a colony at the Cambridge Department of Zoology and were raised individually to maintain phonotactic responsiveness³⁰. After the final moult, a metal pin (32 mg; cricket, 1.2 g) was attached vertically with wax to the first abdominal tergites, close to the animal's centre of gravity. The cerci were covered with wax to prevent the animals responding to air currents. The crickets were first positioned in natural walking posture on the top of the trackball and then the attached pins were clamped in the needle holder. While the animals rotated the trackball with their legs their body position and orientation remained constant and thus they were always exposed to identical acoustic conditions. This is an advantage over closed-loop experiments in which movements of the animal will change the impact of the sound stimuli. All experiments were performed in the dark at a temperature of 24–28 °C.

Trackball recordings

The trackball (Rohacell, diameter of 56.5 mm, 3.0 g) fitted into an acrylic half-sphere with 24 evenly spaced holes mounted into a cylinder (Fig. 1a). Air was passed through the holes of the half-sphere so that the trackball was gently lifted and was free to rotate with minimal friction (Fig. 1a). An optical sensor (Agilent ADNS-2051, 2-D Optical Mouse Sensor) was aligned opposite the south pole of the trackball. Looking through the transparent acrylic sphere it monitored any movements of the trackball in the forward-backward (X) and lateral left-right (Y) direction on two separate data channels. Any trackball movement of 127 μm along a measuring axis produced a short (150 μs) coding pulse in the corresponding data channel, with the sign of the pulse coding the movement direction. The system performed linearly up to speeds of 38 cm s^{-1} in either direction. We did not bin the coding pulses, to maintain the maximum sensitivity of the system. The coding pulses for both movement components and the envelope of the sound stimuli were sampled online at 10 kHz per channel using an A/D board (National Instruments PCI-Mio 16-E-4) controlled by software programmed in LabView 5.01. Data were stored on the hard disk of a PC for off-line analysis. From the pulses coding the lateral movements of the trackball, we calculated the actual lateral steering velocity by which an animal steered to the left or right and the overall lateral deviation of the animal from a straight path. From both movement components we determined the translation velocity of the animal and the path length covered in any test series.

Acoustic stimulation

Sound stimuli were generated with Cool Edit 2000 and were presented by standard PC audio boards through two active speakers (Sony SRS A57). These were positioned in front of the cricket at a distance of 66 cm and at an angle of 45° to the left and right of the animal's longitudinal axis. Sound intensities were adjusted to 75 dB SPL relative to 10⁻⁵ N m⁻² at the position of the cricket and were measured with a Brüel and Kjær free field microphone (type 4191) and measuring amplifier (type 2610). The standard sound pattern had a frequency of 4.8 kHz, a pulse duration of 21 ms, an interval of 21 ms, six pulses per chirp corresponding to a chirp duration of 250 ms and a chirp period of 500 ms. To analyse phonotactic walking, each sound pattern was generally presented from the left and right speaker for a minimum of 30 s duration. The envelope of the sound pattern was calculated on-line using an RMS (root mean square) chip (Analogue Devices, type 637) set to a time constant of 0.5 ms. The experiments were performed inside a sound-proofed chamber. The noise level measured at the top of the trackball was 38 dB SPL (band-pass filter 200–200,000 Hz) with the trackball air supply turned on.

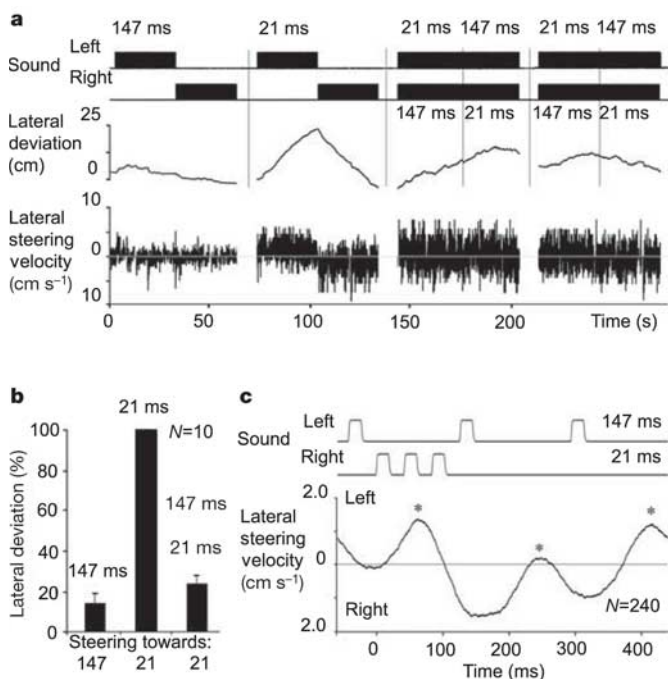


Figure 4 Cricket orientation to patterns of different attractiveness. **a**, A pattern with three sound pulses with intervals of 147 ms elicited only weak orientation towards the sound (first trace). Crickets clearly steered towards a pattern of three pulses with intervals of 21 ms (second trace). Steering towards the 21-ms-interval pattern strongly decreased when both patterns were presented simultaneously (third and fourth traces). **b**, Lateral deviation of ten females to the 147-ms-interval pattern, 21-ms-interval pattern and to the 21-ms-interval pattern when both patterns were presented simultaneously. **c**, An average of the lateral steering velocity ($N = 240$ chirps) shows that the animals steered towards sound pulses from both sides. Error bars in **b** indicate s.e.m.

Recording of leg movements and flight steering

Stance–swing movements of a front leg during walking (Fig. 1b) or the lateral steering movements of the abdomen²⁰ during flight (Fig. 2e) were measured with an optoelectronic camera²⁷. A piece of reflective disc (3M, Scotchlite 7610) was attached to the femur or abdomen and illuminated. The reflected light was picked up by a position-sensitive photodiode and indicated the position of the body part. For measurements during flight, crickets were tethered by their pins in front of a constant air stream to elicit flying. Conditions of acoustic stimulation were otherwise identical to the trackball experiments.

Received 13 May; accepted 25 June 2004; doi:10.1038/nature02787.

1. Webster, D. B., Fay, R. R. & Popper, A. N. *The Evolutionary Biology of Hearing* (Springer, Berlin, New York, 1992).
2. Pollack, G. S. Analysis of temporal patterns of communication signals. *Curr. Opin. Neurobiol.* **11**, 734–738 (2001).
3. Balakrishnan, R., von Helversen, D. & von Helversen, O. Song pattern recognition in the grasshopper *Chorthippus biguttulus*: the mechanism of syllable onset and offset detection. *J. Comp. Physiol. A* **187**, 255–264 (2001).
4. Weber, T. & Thorson, J. In *Crickets: Behaviour and Neurobiology* (eds Huber, F., Moore, T. E. & Lohr, W.) 310–339 (Cornell Univ. Press, Ithaca, London, 1989).
5. Hoy, R. R. Acoustic communication in crickets: a model system of feature detection. *Fed. Proc.* **37**, 2316–2323 (1978).
6. Schildberger, K. Temporal selectivity of identified auditory neurons in the cricket brain. *J. Comp. Physiol. A* **155**, 171–185 (1984).
7. Hennig, R. M. Acoustic feature extraction by cross correlation in crickets? *J. Comp. Physiol. A* **189**, 589–598 (2003).
8. Pollack, G. S. Discrimination of calling song models by the cricket, *Teleogryllus oceanicus*: the influence of sound direction on neural encoding of the stimulus temporal pattern and on phonotactic behaviour. *J. Comp. Physiol. A* **158**, 549–561 (1986).
9. Stabel, J., Wendler, G. & Scharstein, H. Cricket phonotaxis: localization depends on recognition of the calling song pattern. *J. Comp. Physiol. A* **165**, 165–177 (1989).
10. von Helversen, D. & von Helversen, O. Acoustic pattern recognition and orientation in orthopteran insects: parallel or serial processing? *J. Comp. Physiol. A* **177**, 767–774 (1995).
11. Webb, B. Robots in invertebrate neuroscience. *Nature* **417**, 359–363 (2002).
12. Webb, B. & Scutt, T. A simple latency-dependent spiking-neuron model of cricket phonotaxis. *Biol. Cybern.* **82**, 247–269 (2000).
13. Arkin, R. C. *Behaviour-Based Robotics* (MIT, Cambridge, London, 1998).
14. Gerhard, H. C. & Huber, F. *Acoustic Communication in Insects and Anurans* (Univ. of Chicago Press, Chicago and London, 2002).
15. Huber, F. Cricket neuroethology: neuronal basis of intraspecific acoustic communication. *Adv. Study Behav.* **19**, 299–356 (1990).
16. Pollack, G. S. Who, what, where? Recognition and localization of acoustic signals by insects. *Curr. Opin. Neurobiol.* **10**, 763–767 (2000).
17. Popov, A. V. & Shuvalov, V. F. Phonotactic behaviour of crickets. *J. Comp. Physiol. A* **119**, 111–126 (1977).
18. Doherty, J. A. Temperature coupling and trade-off phenomena in the acoustic communication system of the cricket, *Gryllus bimaculatus* de Geer (Gryllidae). *J. Exp. Biol.* **114**, 17–35 (1985).
19. Murphey, R. K. & Zaretsky, M. D. Orientation to calling song by female crickets, *Scasipedeus marginatus* (Gryllidae). *J. Exp. Biol.* **56**, 335–352 (1972).
20. Ulagarai, S. M. & Walker, T. J. Phonotaxis of crickets in flight: attraction of male and female crickets to male calling songs. *Science* **182**, 1278–1279 (1973).
21. Moiseff, A., Pollack, G. S. & Hoy, R. R. Steering responses of flying crickets to sound and ultrasound: mate attraction and predator avoidance. *Proc. Natl Acad. Sci. USA* **75**, 4052–4056 (1978).
22. Schildberger, K., Huber, F. & Wohlers, D. W. In *Crickets: Behaviour and Neurobiology* (eds Huber, F., Moore, T. E. & Lohr, W.) 423–458 (Cornell Univ. Press, Ithaca, London, 1989).
23. Weber, T., Thorson, J. & Huber, F. Auditory behaviour of the cricket. I: Dynamics of compensated walking and discrimination paradigms on the Kramer treadmill. *J. Comp. Physiol. A* **141**, 215–232 (1981).
24. Schmitz, B., Scharstein, H. & Wendler, G. Phonotaxis in *Gryllus campestris* L. (Orthoptera, Gryllidae). I: Mechanism of acoustic orientation in intact female crickets. *J. Comp. Physiol. A* **148**, 431–444 (1982).
25. Doherty, J. A. Song recognition and localization in the phonotaxis behavior of the field cricket, *Gryllus bimaculatus* (Orthoptera: Gryllidae). *J. Comp. Physiol. A* **168**, 213–222 (1991).
26. Weber, T. & Thorson, J. Auditory behaviour of the cricket. IV: Interaction of direction of tracking with perceived split-song paradigms. *J. Comp. Physiol. A* **163**, 13–22 (1988).
27. Hedwig, B. A highly sensitive opto-electronic system for the measurement of movements. *J. Neurosci. Methods* **100**, 165–171 (2000).
28. Pollack, G. S. & Hoy, R. R. Phonotaxis in flying crickets: neural correlates. *J. Insect Physiol.* **27**, 41–45 (1981).
29. Nabatyan, A., Poulet, J. F. A., de Polavieja, G. G. & Hedwig, B. Temporal pattern recognition based on instantaneous spike rate coding in a simple auditory system. *J. Neurophysiol.* **90**, 2484–2493 (2003).
30. Cade, W. H. Effect of male deprivation on female phonotaxis in field crickets (Orthoptera: Gryllidae; *Gryllus*). *Can. Entomol.* **111**, 741–744 (1979).

Acknowledgements We thank our Cambridge and Edinburgh colleagues for comments on the manuscript. The BBSRC and the Royal Society supported the project. We are grateful to M. Knepper and P. Williams for the development of software and hardware and to Röhm GmbH for providing Rohacell.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.H. (bh202@cam.ac.uk) or J.F.A.P. (jlap2@cam.ac.uk).

MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode

Sarah Chang¹, Robert J. Johnston Jr¹, Christian Frøkjær-Jensen², Shawn Lockery² & Oliver Hobert¹

¹Department of Biochemistry and Molecular Biophysics, Center for Neurobiology and Behavior, Columbia University Medical Center, 701 W. 168th Street, New York 10032, USA

²Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, USA

Animal microRNAs (miRNAs) are gene regulatory factors that prevent the expression of specific messenger RNA targets by binding to their 3' untranslated region^{1–3}. The *Caenorhabditis elegans* *lsy-6* miRNA (for lateral symmetry defective) is required for the left/right asymmetric expression of guanylyl cyclase (*gcy*) genes in two chemosensory neurons termed ASE left (ASEL) and ASE right (ASER)^{4,5}. The asymmetric expression of these putative chemoreceptors in turn correlates with the functional lateralization of the ASE neurons⁶. Here we find that a mutation in the *die-1* zinc-finger transcription factor disrupts both the chemosensory laterality and left/right asymmetric expression of chemoreceptor genes in the ASE neurons. *die-1* controls chemosensory laterality by activating the expression of *lsy-6* specifically in ASEL, but not in ASER, where *die-1* expression is downregulated through two sites in its 3' untranslated region. These two sites are complementary to *mir-273*, a previously uncharacterized miRNA, whose expression is strongly biased towards ASER. Forced bilateral expression of *mir-273* in ASEL and ASER causes a loss of asymmetric *die-1* expression and ASE laterality. Thus, an inverse distribution of two sequentially acting miRNAs in two bilaterally symmetric neurons controls laterality of the nematode chemosensory system.

Although miRNAs are abundant in animal genomes, the biological contexts and pathways in which miRNAs operate are only beginning to be explored³. The *C. elegans* *lsy-6* miRNA functions in a poorly understood developmental context, the generation of neuronal diversity along the left/right axis of an animal^{4,5}. *lsy-6* superimposes a left/right asymmetric expression profile of putative chemosensory receptors, encoded by the *gcy* genes, onto the bilaterally symmetric differentiation program of two chemosensory neurons, ASEL and ASER^{4,5}. Asymmetric *gcy* chemoreceptor expression in turn correlates with the left/right asymmetric chemosensory capacities of ASEL and ASER⁶. An essential prerequisite for ASEL/R laterality is the restriction of *lsy-6* expression to the ASEL neuron⁵.

To gain a better mechanistic understanding of *lsy-6*-mediated lateralization of the ASE neurons, we conducted genetic screens for mutants that show defects in asymmetric expression of ASE-specific putative *gcy* chemoreceptors (see Supplementary Information)⁷. One of the alleles retrieved from this screen, *ot26*, showed a 100% penetrant *lsy* phenotype in adult animals; both ASE cells expressed the normally ASER-specific *gcy-5* gene and concomitantly lost the expression of the normally ASEL-specific *gcy-7* gene (Fig. 1a, b). The expression of several cell-fate markers that label bilaterally symmetric aspects of ASEL/R differentiation were unaffected (data not shown).

Further analysis of *ot26* mutants showed that the laterality defects at the genetic level were tightly correlated with defects at the behavioural level. Previous studies of chemotaxis behaviour in worms, in which either the left or right ASE neuron was killed,

indicated that ASEL and ASER are functionally distinct; ASEL responds strongly to Na^+ ions, and not at all to Cl^- ions, whereas ASER responds strongly to Cl^- ions, but only weakly to Na^+ ions (Fig. 1c; wild-type-model)⁶. Given that in *ot26* mutants, ASEL seems to adopt an ASER-specific chemoreceptor expression profile (Fig. 1b), *ot26* mutants might be expected to behave as though they have two ASER neurons (Fig. 1c, 2-ASER model). If so, chemotaxis to Na^+ ions, which are a weaker stimulus for ASER, should be impaired, whereas chemotaxis to Cl^- ions should be intact or even enhanced. As predicted by the 2-ASER model, *ot26* mutants were indeed defective in locating the peak of a gradient of Na^+ ions, especially in the presence of a high, uniform concentration of Cl^- ions (Fig. 1d, e). Furthermore, *ot26* mutants were better at finding the peak of a gradient of Cl^- ions in the presence of a high, uniform concentration of Na^+ ions (Fig. 1f). Therefore, the *ot26* allele affects a gene that is necessary for forming distinct neuronal representations of chemosensory inputs.

Single-nucleotide-polymorphism mapping, complementation testing, transformation rescue and allele sequencing showed that *ot26* is an allele of the *die-1* gene (Supplementary Fig. 1a). The *die-1* gene encodes a C2H2 zinc-finger transcription factor previously implicated in hypodermal patterning⁸. Other *die-1* alleles show similar lateralization defects (Supplementary Fig. 1a). A *die-1^{resc}::gfp* reporter gene, which rescues the *lsy* defect (Supplementary Fig. 1b), shows a dynamic expression profile with early hypodermal expression ceasing at gastrulation and then reappearing in late embryos in several putative head neurons⁸. Using an

ASEL-neuron-expressed red-fluorescent-protein reporter (*ceh-36^{prom}::rfp*), we identified two of these neurons as ASEL and ASER. However, expression is strongly biased towards ASEL (Fig. 2a). We corroborated the autonomous function of *die-1* in the mature ASEL neurons by expressing the *die-1* complementary DNA under the control of the promoter of the ASEL-specific *gcy-7* gene. When expressed in a *die-1* mutant background, three of four lines showed rescue of the *lsy* phenotype (Supplementary Fig. 1b).

We previously described a pathway consisting of several gene-regulatory factors that controls left/right asymmetric *gcy* gene expression in ASEL/R (summarized in Fig. 4)⁷. Placing *die-1* into this pathway, we found a completely penetrant loss of *lim-6* homeobox gene expression in adult *die-1(ot26)* mutant animals (Supplementary Fig. 2). Because correct *lim-6* expression requires the tightly regulated balance of the *ceh-36* activator homeobox gene and the *cog-1* repressor homeobox gene (Fig. 4)⁷, a loss of *lim-6* expression is consistent with either a decrease in the expression of the *ceh-36* activator or an increase in the expression of the *cog-1* repressor. We found that loss of *die-1* had no effect on *ceh-36* expression ($n = 52$), but instead caused a strong de-repression of *cog-1::gfp* expression in ASEL (Fig. 2b). Furthermore, expressing *die-1* ectopically in the ASER neuron demonstrated that *die-1* is not only necessary but also sufficient to repress *cog-1* expression (Fig. 2b).

One possible explanation for the ability of *die-1* to repress *cog-1* expression is that *die-1* regulates expression of *lsy-6*, a miRNA that we previously have shown to inhibit *cog-1* expression through

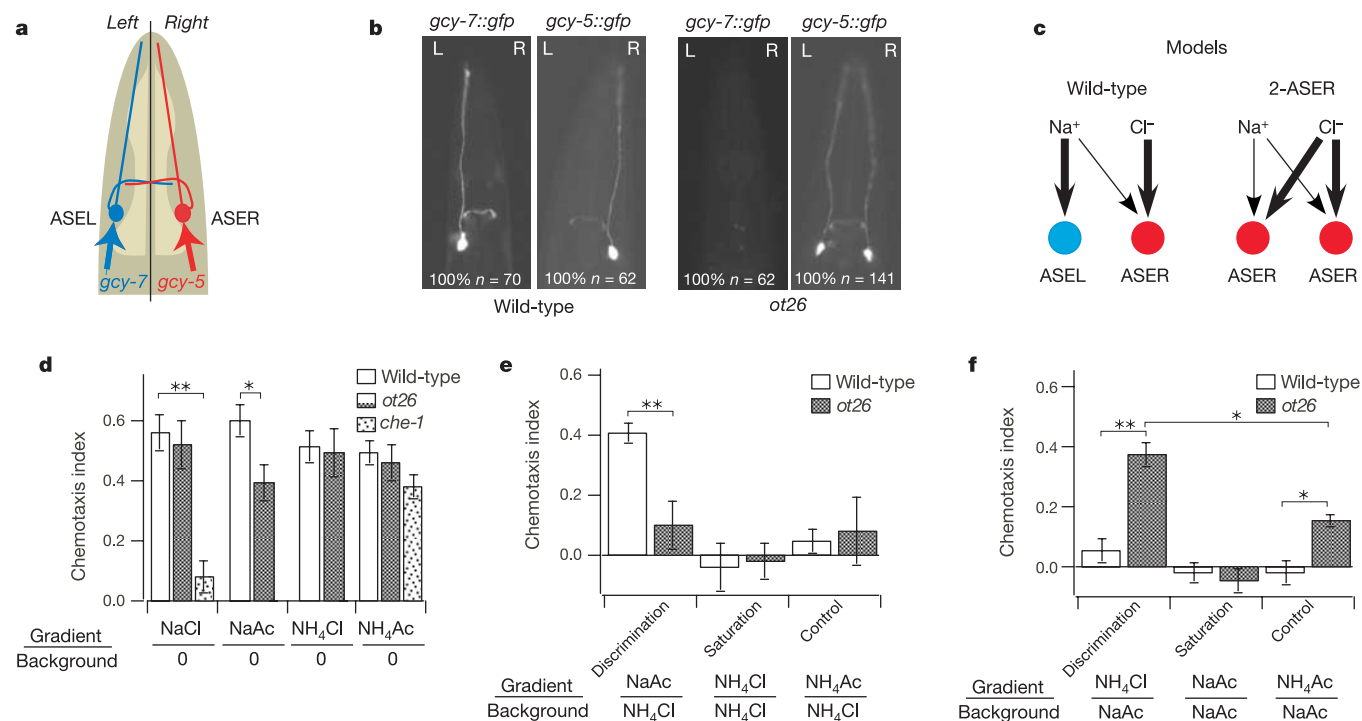


Figure 1 Laterality of ASE neurons is disrupted in *ot26* mutant animals. **a**, Schematic structure of ASEL/R and *gcy* gene expression profiles¹⁶. **b**, Transgenic adult animals expressing *gcy-5^{prom}::gfp* (*ntl5*) or *gcy-7^{prom}::gfp* (*otl3*) in wild-type or *ot26* mutant backgrounds. **c**, Models of chemosensory processing in ASEL/R neurons of wild-type and *ot26* (2-ASER) animals. Arrow thickness represents relative sensitivity to the indicated ion, inferred from laser-ablation experiments⁶. Na^+ versus Cl^- sensitivity in the 2-ASER model predicts reduced Na^+ chemotaxis and normal or enhanced Cl^- chemotaxis for these ions tested alone (**d**), defective Na^+ versus Cl^- discrimination (**e**, 'Discrimination'), and normal or enhanced Cl^- versus Na^+ discrimination (**f**, 'Discrimination'). **d–f**, Average chemotaxis index of wild-type and *ot26* animals in chemotaxis assays and discrimination tests. The composition of the chemical gradient that was superimposed upon the

background stimulus (if any) is shown below each pair of bars. Peak concentration of the gradients was ~ 10 mM; background concentrations were 100 mM. Bars are mean $\pm$ s.e.; three to nine assays per datum (bar) with $N > 100$ animals per assay. Asterisks indicate significance at $P < 0.05$ (*) and $P < 0.01$ (**). All strains contained the *ntl5* (*gcy-5^{prom}::gfp*) transgene in the background. For counter-ions (NH_4 and Ac) and controls (*che-1*) see Supplementary Information. Relative chemotaxis performance of wild-type worms and *ot26* mutants was consistent with the predictions in **c**. The enhanced response of *ot26* mutants to an NH_4Ac gradient (**f**, 'Control') is significantly weaker than the enhanced NH_4Cl response (**f**, 'Discrimination'), confirming the Cl^- specificity of the latter.

binding to its 3' untranslated region (UTR)⁴. In support of this, we found that *lsy-6* expression in ASEL is abolished in a *die-1* mutant background (Fig. 2c). In other neurons, *die-1* has no effect on *lsy-6* expression, illustrating that the promoter of the *lsy-6* locus samples distinct transcriptional regulatory inputs in different cell types. In the context of the ASE neurons, *die-1* is not only necessary but also sufficient to control *lsy-6* expression, because ectopic expression of *die-1* in ASER results in ectopic *lsy-6^{prom}::gfp* expression (Fig. 2c). We confirmed that *die-1* acts through *lsy-6* to repress *cog-1* expression by demonstrating that transgenic animals that express *die-1* in ASER are not able to turn off *cog-1* expression in a *lsy-6* null mutant background (Fig. 2b).

We next investigated how *die-1* and hence *lsy-6* activation is spatially biased towards ASEL. Because the inverse asymmetry of *cog-1* expression in this regulatory cascade (ASER > ASEL expression) is controlled by *lsy-6* binding to the *cog-1* 3' UTR, we asked whether *die-1* expression is also controlled by its 3' UTR. To test this idea, we constructed a 'sensor gene' in which *gfp* transcripts were produced under the control of the *ceh-36* promoter in both ASEL and ASER. When *gfp* transcripts are fused to a heterologous 3' UTR, GFP expression is observed in ASEL and ASER at a

comparable level (Fig. 3a)^{4,7}. If the transcript is fused to the 3' UTR of the *cog-1* gene, which is responsive to the ASEL-expressed *lsy-6* miRNA, downregulation of GFP expression is observed in ASEL⁴. If the *gfp* transcript is fused to the 3' UTR of the *die-1* gene, GFP expression is significantly downregulated in ASER (Fig. 3a). The 3' UTR of *die-1* therefore contains elements that are subject to left/right-specific control of expression, providing a mirror image of control-of-expression of the *cog-1* 3' UTR.

We next sought to identify a miRNA that may control *die-1* expression through binding to the 3' UTR of *die-1*. Using an approach conceptually similar to 'phylogenetic footprinting' (traditionally used to identify transcription-factor binding sites in putative transcriptional regulatory regions⁹), we compared the 3' UTRs of the *die-1* gene from *C. elegans* with those from the related nematode, *Caenorhabditis briggsae*. We found three phylogenetic footprints with consecutive stretches of >10-nucleotide sequence identity in the respective *die-1* 3' UTRs (Supplementary Fig. 3a). When these were translated into RNA, we noted that two of the sites each showed complementarity to a hypothetical stretch of six nucleotides, 3'-UG(C/U)(C/U)(C/U)G-5' (Supplementary Fig. 3a). We inferred that this stretch of six nucleotides is a part of

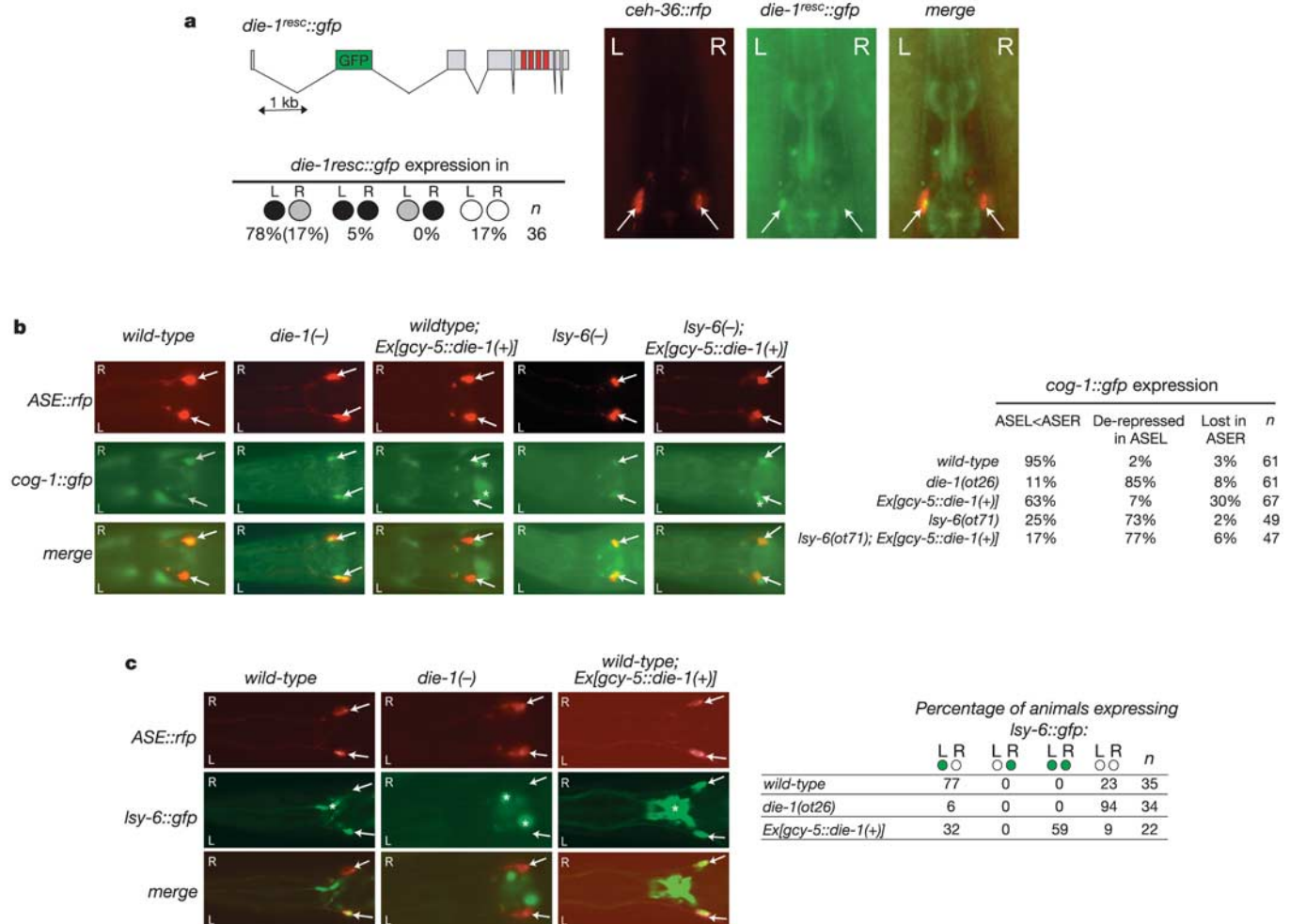


Figure 2 The zinc-finger transcription factor, DIE-1, is predominantly expressed in ASEL and affects *lsy-6* expression. **a**, Expression pattern of a chromosomally integrated *die-1^{resc}::gfp* reporter construct (*otIs159*). ASE cells were identified with the *otIs151* (*ceh36^{prom}::rfp*) array. See Supplementary Information for more details on expression patterns and constructs. Circles indicate expression levels. Number in parenthesis indicates expression exclusively in ASEL. **b**, *cog-1* expression in various genetic backgrounds, monitored with the *syIs63* array, which contains the complete *cog-1*

locus¹⁷. ASEL/R (shown by arrows) were identified with a *ceh-36^{prom}::rfp* transgene, which is also expressed in the AWC neurons. Asterisks point to other *cog-1::gfp* expressing cells. The tabular data quantify the expression levels (some animals appear in more than one category). **c**, Expression of *lsy-6^{prom}::gfp* (*otEx1403*) in wild-type (left panels), in *die-1(ot26)* mutant animals (middle panels) and transgenic wild-type animals that express *die-1* in ASER (*otEx1192* array) (left panels). Asterisks point to other *lsy-6^{prom}::gfp*-expressing cells.

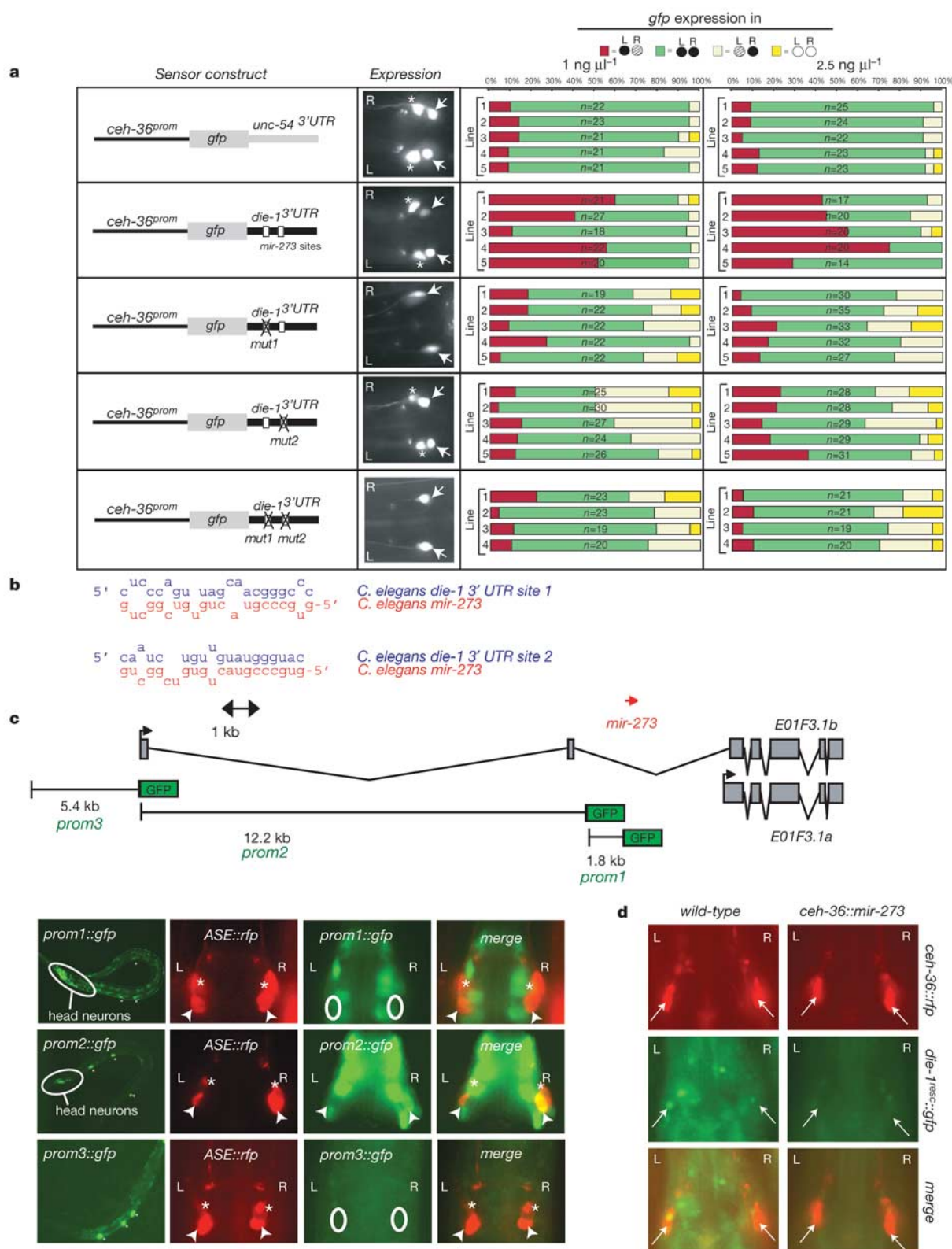


Figure 3 *die-1* regulation by the *mir-273* miRNA. **a**, Expression of *die-1* 3' UTR sensor constructs in ASEL and ASER. In 'mut1' the 'acgggc' core of the *mir-273* complementary site is changed to 'guaacu'; in 'mut2' the 'gggua' core is changed to 'cuacg'. Sensor constructs were injected at two different concentrations as indicated. **b**, Complementarity of *die-1* 3' UTR sites and *mir-273*. **c**, *mir-273* expression pattern deduced from reporter gene fusions. *prom1* and *prom2* are exclusively expressed in head neurons; *prom2* shows biased expression in the ASER neuron (for quantification, see

Supplementary Table 1). Several transgenic lines show similar expression patterns. *ceh-36^{prom1}::rfp* (*otIs151*) identifies ASEL and ASER. Asterisks in left panels indicate cells that fluoresce owing to the injection marker used (*unc-122^{prom1}::gfp*); asterisks in other panels indicate AWC neurons. **d**, Bilateral expression of *mir-273* (*ceh-36^{prom1}::mir-273*; *otEx1705*) disrupts *die-1* expression, assayed with the *otIs159* (*die-1^{resc}::gfp*) transgene (for quantification, see Supplementary Table 2).

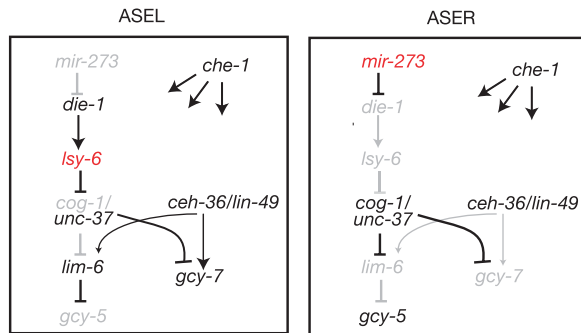


Figure 4 Gene regulatory cascade in ASEL and ASER. Genetic data from this paper and refs 4 and 7 are summarized. miRNAs acting in this cascade are indicated in red.

a miRNA that exerts control over the 3' UTR. We examined the sequences of all predicted miRNAs in the *C. elegans* genome (<http://www.sanger.ac.uk/Software/Rfam/mirna/>) and found a match for this sequence in the functionally uncharacterized *mir-273* miRNA (Supplementary Fig. 3b)¹⁰. The complementarity of *mir-273* to the two phylogenetic footprints in the *die-1* 3' UTR extends to a total of 19–21 nucleotides (Fig. 3b), a common length of miRNA/target duplexes^{1–3,11}. As expected from the conserved nature of the six-nucleotide sequence with which we conducted the search, this sequence lies at the 5' end of *mir-273* (Fig. 3b), where most known miRNAs show the highest degree of sequence identity to their target sites¹¹. To investigate whether the 3' UTR-mediated downregulation of *die-1* expression depends on the two *mir-273*-complementary sites, we mutated each of the sites in the sensor *gfp* construct. Each of the mutant sensor constructs failed to be downregulated in ASER (Fig. 3a). The requirement for each complementary site is consistent with the previously observed cooperativity of miRNA action¹².

The downregulation of the *die-1* 3' UTR in a manner that is (1) dependent on the *mir-273* complementary sites, and (2) specific to the ASER cell, suggests that *mir-273* activity is biased towards ASER, possibly through left/right differential gene expression. Using *gfp* reporter constructs that sample potential transcriptional *cis*-regulatory inputs into the *mir-273* locus, we inferred that *mir-273* is expressed in several bilaterally symmetric pairs of head neurons (Fig. 3c). Left/right asymmetric expression was indeed observed in the ASE neurons, with expression in ASER being significantly higher than in ASEL (Fig. 3c; Supplementary Table 1). If the asymmetric distribution of *mir-273* was functionally relevant, we would expect that forced, symmetric expression of *mir-273* in both ASEL and ASER would repress *die-1* expression and hence disrupt ASE laterality. Indeed, transgenic animals that express *mir-273* from the bilateral *ceh-36* promoter (*ceh-36^{prom}::mir-273*) not only show a downregulation of *die-1^{resc}::gfp* expression (Fig. 3d), but also show the 2-ASER chemoreceptor profile that is characteristic of the *die-1* loss-of-function phenotype (Supplementary Fig. 4). Diametrically opposite effects (2-ASEL phenotype) can be observed if the normally left-side-specific miRNA *lsy-6* is expressed under control of the bilateral *ceh-36* promoter⁴.

Starting with the identification of a mutant strain that shows laterality defects both in gene expression patterns and in functional capacities, we have described a critical transcription factor component of a gene regulatory cascade that controls left/right asymmetry in the chemosensory system of *C. elegans*. This cascade uses several fundamental principles of gene regulation, one being the use of a series of sequential repressive interactions¹³, another being the extensive use of miRNA-mediated control of gene expression (Fig. 4). The spatial restriction of the expression of one of the miRNAs involved in this pathway, *lsy-6*, is controlled through another spatially controlled miRNA, *mir-273*. This regulatory effect

is mediated by the *die-1* zinc-finger transcription factor, the first transcription factor found to confer spatial control of animal miRNA gene expression. Therefore, ASEL and ASER show reciprocal miRNA expression profiles, an essential requirement for determining laterality of the chemosensory system. □

Methods

Strains, genetic screening and mapping

A detailed list of strains and genetic approaches can be found in the Supplementary Information.

DNA constructs and transgenes

A *die-1*-rescuing *gfp* plasmid (pPH12, '*die-1^{resc}::gfp*') was provided by J. Hardin (see Supplementary Information)⁸. Polymerase chain reaction (PCR) fusion¹⁴ was used to generate *mir-273* promoter fusion constructs and 3' UTR sensor constructs. For the latter, five kilobases (kb) of the *ceh-36* promoter were PCR-fused to *gfp* coding sequences and then, in a second PCR fusion reaction, fused to an *unc-54* 3' UTR or the *die-1* 3' UTR-PCR amplicon. To generate the mutated sensor constructs, the *die-1* 3' UTR was amplified with a total of four PCR primers that harbour the mutated *mir-273* complementary sites. To generate *gcy-5^{prom}::die-1* and *gcy-7^{prom}::die-1*, *gfp* from *gcy-5Hind::gfp* and *gcy-7Hind::gfp* expression constructs⁷ was replaced with a *die-1* cDNA (provided by J. Hardin). PCR fusion was used to generate *mir-273* misexpression constructs, using the 5-kb *ceh-36* promoter and the predicted 94-base pair (bp) *mir-273* hairpin. See Supplementary Information for PCR primer sequences used for PCR fusions. Transgenic lines were usually generated in a wild-type (strain N2) background using *unc-122^{prom}::gfp* (50 ng μ l⁻¹) or *rol-6* (100 ng μ l⁻¹) as injection markers.

Chemotaxis assays

Radial gradient assays were based on previously described assays¹⁵. A detailed description can be found in the Supplementary Information.

Received 17 March; accepted 14 June 2004; doi:10.1038/nature02752.

- Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
- Wightman, B., Ha, I. & Ruvkun, G. Post-transcriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell* **75**, 855–862 (1993).
- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
- Johnston, R. J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845–849 (2003).
- Hobert, O., Johnston, R. J. & Chang, S. Left-right asymmetry in the nervous system: the *Caenorhabditis elegans* model. *Nature Rev. Neurosci.* **3**, 629–640 (2002).
- Pierce-Shimomura, J. T., Faumont, S., Gaston, M. R., Pearson, B. J. & Lockery, S. R. The homeobox gene *lin-6* is required for distinct chemosensory representations in *C. elegans*. *Nature* **410**, 694–698 (2001).
- Chang, S., Johnston, R. J. & Hobert, O. A transcriptional regulatory cascade that controls left/right asymmetry in chemosensory neurons of *C. elegans*. *Genes Dev.* **17**, 2123–2137 (2003).
- Heid, P. J. et al. The zinc-finger protein DIE-1 is required for late events during epithelial cell rearrangement in *C. elegans*. *Dev. Biol.* **236**, 165–180 (2001).
- Bulyk, M. L. Computational prediction of transcription-factor binding site locations. *Genome Biol.* **5**, 201 (2003).
- Grad, Y. et al. Computational and experimental identification of *C. elegans* microRNAs. *Mol. Cell.* **11**, 1253–1263 (2003).
- Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. Identification of *Drosophila* microRNA targets. *PLoS Biol.* **1**, 1–13 (2003).
- Doench, J. G., Petersen, C. P. & Sharp, P. A. siRNAs can function as miRNAs. *Genes Dev.* **17**, 438–442 (2003).
- Muhr, J., Andersson, E., Persson, M., Jessell, T. M. & Ericson, J. Groucho-mediated transcriptional repression establishes progenitor cell pattern and neuronal fate in the ventral neural tube. *Cell* **104**, 861–873 (2001).
- Hobert, O. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *Biotechniques* **32**, 728–730 (2002).
- Bargmann, C. I. & Horvitz, H. R. Chemosensory neurons with overlapping functions direct chemotaxis to multiple chemicals in *C. elegans*. *Neuron* **7**, 729–742 (1991).
- Yu, S., Avery, L., Baude, E. & Garbers, D. L. Guanylyl cyclase expression in specific sensory neurons: a new family of chemosensory receptors. *Proc. Natl Acad. Sci. USA* **94**, 3384–3387 (1997).
- Palmer, R. E., Inoue, T., Sherwood, D. R., Jiang, L. I. & Sternberg, P. W. *Caenorhabditis elegans* *cog-1* locus encodes GTX/Nkx6.1 homeodomain proteins and regulates multiple aspects of reproductive system development. *Dev. Biol.* **252**, 202–213 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Antonio for the *ot100* allele, J. Hardin for *die-1* reagents, Q. Chen and S. Narula for expert technical assistance, J. Tien for help with mapping *ot26*, I. Greenwald for advice, L. Johnston, A. Lanjuin, P. Sengupta, I. Greenwald and F. Slack for reading the manuscript. S.C. was funded by an NIH predoctoral fellowship, R.J.J. by an NSF predoctoral fellowship, and C.F.-J. by a American Heart Association predoctoral fellowship. O.H. was supported by grants from the NIH, the McKnight Foundation and the Irma T. Hirsch Trust.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to O.H. (or38@columbia.edu)

Reaction mechanism determines NMDA receptor response to repetitive stimulation

Gabriela Popescu¹, Antoine Robert², James R. Howe² & Anthony Auerbach¹

¹University at Buffalo, Department of Physiology and Biophysics and Center for Single Molecule Biophysics, Buffalo, New York 14214, USA

²Department of Pharmacology, Yale University School of Medicine, New Haven, Connecticut 06520, USA

At central excitatory synapses, *N*-methyl-D-aspartate (NMDA) receptors, which have a high affinity for glutamate¹, produce a slowly rising synaptic current in response to a single transmitter pulse and an additional current after a second, closely timed stimulus². Here we show, by examining the kinetics of transmitter binding and channel gating in single-channel currents from recombinant NR1/NR2A receptors, that the synaptic response to trains of impulses is determined by the molecular reaction mechanism of the receptor. The rate constants estimated for the activation reaction predict that, after binding neurotransmitter, receptors hesitate for ~4 ms in a closed high-affinity conformation before they either proceed towards opening or release neurotransmitter, with about equal probabilities. Because only about half of the initially fully occupied receptors become active, repetitive stimulation elicits currents with distinct waveforms depending on pulse frequency. This high-affinity/low-efficiency activation mechanism might serve as a link between stimulation frequency and the directionality of the ensuing synaptic plasticity.

Given the high affinity of NMDA receptors (NRs) for glutamate estimated from macroscopic current dose–response profiles (equilibrium dissociation constant $K_d \sim 3 \mu\text{M}$)^{1,3} and the estimated high glutamate concentration in the cleft after the release of a single synaptic vesicle (1–5 mM)^{4,5}, it was widely believed that the NR glutamate-binding sites become saturated after each synaptic vesicle release⁶. This assumption has been called into question by studies indicating that the synaptic glutamate transient might be lower and shorter than previously thought⁷ and by evidence that at some central excitatory synapses a single release event does not elicit a maximal NR current response^{2,8–11}.

Macroscopic current measurements give estimates of apparent affinity that depend inextricably on the gating properties of the channel, whereas single-channel kinetic analyses permit ligand binding and conformational transitions to be determined separately. To estimate the glutamate association and dissociation rate constants, and hence their ratio, K_d , we recorded single-channel currents from cell-attached patches of HEK293 cells transiently expressing rat NR1 and NR2A subunits under experimental conditions that preclude the occupancy of extraneous receptor states such as those resulting from cation binding in the pore¹² or at allosteric extracellular sites on the receptor^{13,14}. Activity was elicited by steady-state saturating concentrations of glycine (100 μM) plus several concentrations of glutamate (20 nM to 2 mM). As reported previously for saturation conditions^{15,16}, individual NRs generated three patterns of activity ('modes') at all glutamate concentrations tested (Fig. 1). We classified periods with homogenous gating kinetics according to channels' mean open times (MOTs), namely high (H, ~33 ms), medium (M, ~11 ms) or low (L, ~3.7 ms) (Supplementary Table S1). Similar results were obtained with several concentrations (0.1–100 μM) of the co-agonist glycine in the presence of 1 mM glutamate (data not shown). A channel's MOT tended to decrease in discrete steps with recording time, as

observed with high concentrations of agonists, with no apparent correlation between agonist concentration and modal behaviour. We conclude that agonist concentration does not affect the modal behaviour of the channel.

To determine the kinetics of glutamate binding to NRs with distinct gating patterns we analysed H, M and L currents separately¹⁷. First, we separated subpopulations of active-time segments (that is, sections devoid of long non-conducting periods associated with desensitization) that had similar MOTs. We focused initially on periods with M-mode kinetics (Fig. 2a, b) because this behaviour was the most common. Within each record, active-time segments with M-mode kinetics were pooled and fitted by models having, as a core, three closed and two open states (3C2O). For activity recorded at glutamate concentrations of less than 10 μM , an additional component was apparent in the closed-interval duration distributions (Fig. 2b). The distribution of open times was unaffected by glutamate concentration¹⁸. The time constant for this additional closed-time component increased with decreasing glutamate concentrations, indicating that it might reflect receptor sojourns in closed mono-liganded (C^M) and unliganded (C^U) states. Visits to such closed states whose lifetimes are dependent on concentration

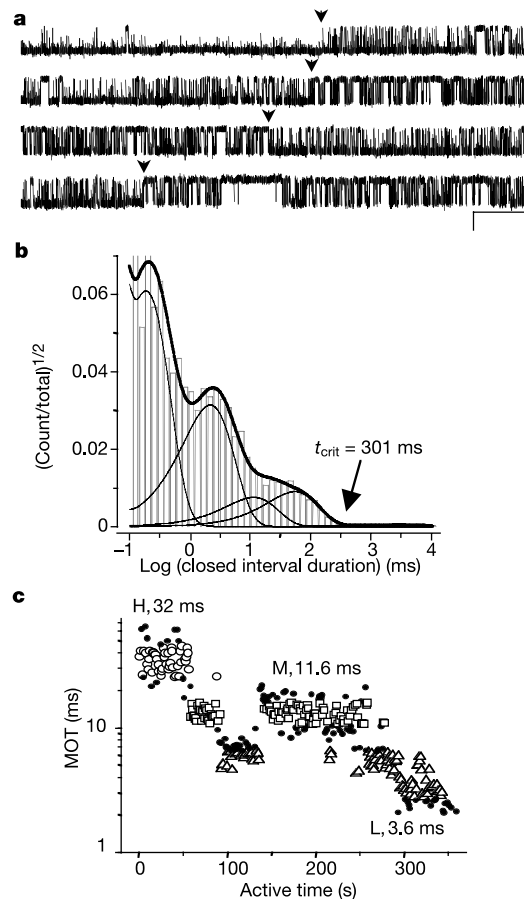


Figure 1 Single-channel NR activity at 5 μM glutamate. **a**, Each current trace (10 s) shows a shift in current pattern (arrow); open is down; scale, 5 pA (vertical) and 1 s (horizontal). **b**, Closed-interval (>75 μs) histogram for the entire record (25,955 intervals). Data were best described by a 5C4O model. Thin lines are the probability density functions for the closed components. The thick line is their sum. Time constants and amplitudes for this record were (in ms (%)) 0.17 (46), 2.1 (35), 10.4 (8.5), 49 (10.3) and 2,799 (0.4). Time constants and amplitudes for the four open states (histogram not shown) were 0.13 (10.8), 3.3 (23), 12.7 (55) and 43.6 (10). Intervals greater than t_{crit} were excluded from further analysis. **c**, Segments (1 s each) of active time segregated into three populations according to their MOT: open circles, high; squares, medium; triangles, low; filled circles, excluded. The mean for each population is indicated.

can be observed in the active-time record as gaps in current that increase in duration with decreasing glutamate concentration (Fig. 2a). These intervals can be detected in the closed-interval distributions when their mean lifetime is longer than that of the longest active-time closed component but shorter than those of the components associated with desensitization (Fig. 2b). Kinetic parameters for NR1/2A receptors gating in each mode are given in Supplementary Table S1.

The relationship between time and rate constants is dependent on the model used for optimisation, therefore, our objective of estimating microscopic rate constants for binding required the selection of an appropriate kinetic scheme. To fit channel activity across glutamate concentrations we extended the core 3C2O model by adding glutamate binding steps as two coupled closed states (C^M and C^U), each having an exit rate constant that was dependent on concentration. We assumed that the binding of glutamate to each of the two binding sites occurred independently and with equal rate constants. There are nine unique ways of incorporating such a pair of coupled states assuming the following: a single gateway state from the core scheme, no cycles in the model, and coupled open states¹⁵ (see Supplementary Fig. S1).

Each of these extended schemes was used to fit jointly active-time intervals from currents with homogenous kinetics recorded under various glutamate concentrations. Although two models emerged as the most likely based on a log-likelihood (LL) criterion, all models predicted similar values for the single-site glutamate association (k_{on}) and dissociation (k_{off}) rate constants (Supplementary Table S2). Figure 2c shows the results for H, M and L currents with the use of the top-ranked scheme in which all the closed states are coupled. The optimal rate constants were essentially the same for all modes: $k_{on} = 2 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$ and $k_{off} = 60 \text{ s}^{-1}$ ($K_d = 3 \mu\text{M}$).

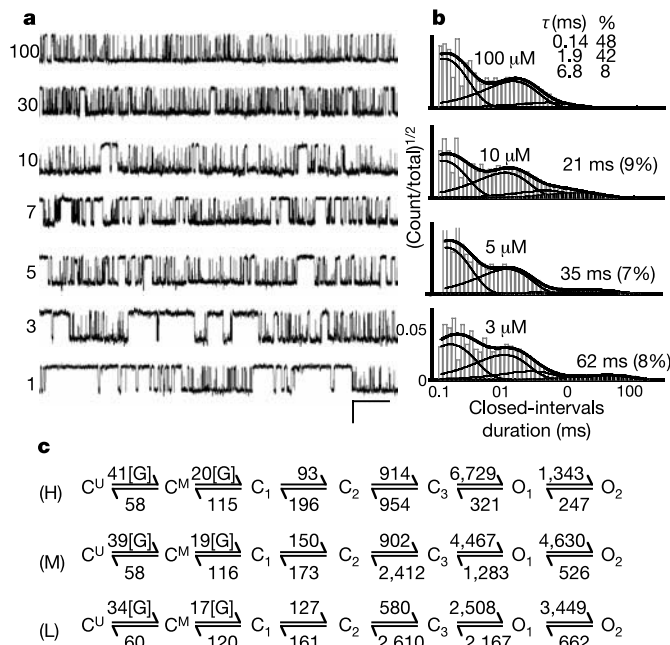


Figure 2 NR activity at several glutamate concentrations. **a**, M-mode active-time currents; concentrations (μM) are indicated at the left. Gaps become longer with decreasing concentration; open is down; scale, 5 pA (vertical) and 250 ms (horizontal). **b**, Each histogram represents M-mode closed intervals from one patch. Superimposed are probability density functions calculated from best-fit models across data pooled from all M currents recorded at a given concentration. Time constants (τ) and amplitudes (%) are given in the top panel for the concentration-independent components and on the other panels for the concentration-dependent component only. **c**, Rate constants ($\mu\text{M}^{-1} \text{ s}^{-1}$ and s^{-1}) optimized for H-, M- and L-gating receptors using the fully sequential model (s.d. < 10%). Microscopic affinity is $3 \mu\text{M}$ per site, regardless of mode.

This value is in agreement with the glutamate affinity of native NRs¹ and recombinant NR1/2A receptors³ estimated from macroscopic measurements. These results confirm that resting NRs have a high affinity for glutamate and show that the microscopic affinity is independent of the mode.

Next we explored the dose-response properties of receptors gating in each mode. Although our results show that the microscopic K_d for glutamate is identical for all modes, the steady-state open probabilities (excluding desensitization) are dependent on the mode (L, 0.4; M, 0.83; H, 0.97). These differences in gating kinetics might influence the apparent macroscopic affinity¹⁹. We used the models and rates illustrated in Fig. 2c to simulate macroscopic NR responses to pulses of glutamate. For all modes the peak current response to a 1-ms pulse of 1 mM glutamate was about half of the single-channel open probability (P_o) at steady state (0.2, 0.45 and 0.41 for L, M and H modes, respectively) and seemed to saturate at 1 mM irrespective of pulse duration (Fig. 3a, b). In contrast, currents elicited by a longer (100-ms) pulse reached peak amplitudes equal to the steady-state open probabilities of single channels for each kinetic mode. Predicted macroscopic dose-response profiles were independent of the mode, with an apparent concentration giving half-maximal response (EC_{50}) of $75 \mu\text{M}$ (Fig. 3c) for short pulses. Receptor dose-response profiles to long pulses were significantly shifted towards lower concentrations and had distinct apparent EC_{50} values for each mode (3.3, 1.1 and $0.7 \mu\text{M}$ for L, M and H, respectively) (Fig. 3b, c). Thus the model predicts correctly that a channel's apparent affinity varies with pulse duration as observed for NRs native to cortical neurons in culture²⁰. The relationship between peak P_o and pulse duration is similar for all modes (Fig. 3d).

Simulated responses from L-mode receptors to a short (1-ms) neurotransmitter pulse reproduced the kinetic characteristics of native and recombinant NR1/2A receptor currents, namely a rise time (10–90%) of ~ 7.5 ms (peak at ~ 13 ms) and a decay time constant of ~ 42 ms (refs 21, 22) (Fig. 3a). Furthermore, the

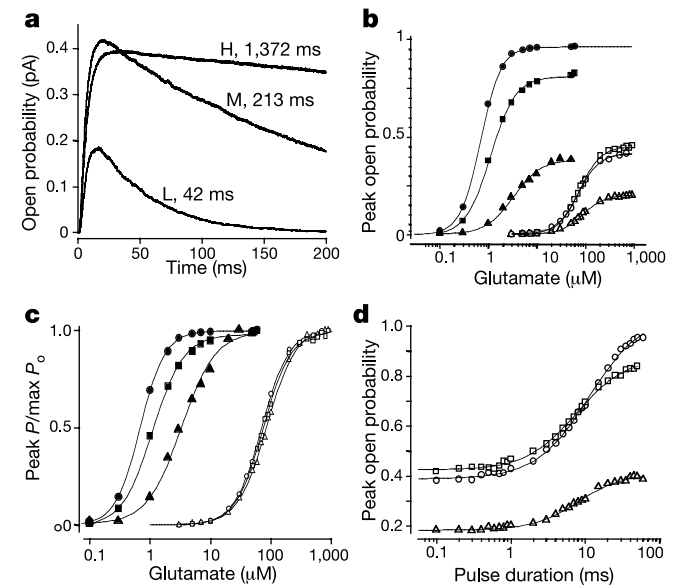


Figure 3 Simulated NR macroscopic responses to a single glutamate pulse. **a**, Responses to a 1-ms pulse of 1 mM glutamate. Peak current amplitude for each mode is about half the maximum P_o (see **d**). Decay time constants obtained by fitting to a single exponential are indicated. **b**, NR dose-response profiles vary with pulse duration. **c**, Normalized dose-response profiles for L-, M- and H-gating receptors. **d**, Peak P_o to a 1-mM pulse varies similarly to pulse duration for either mode and starts at 50% of maximum for short pulses. Symbols in **b–d**: circles, H mode; squares, M mode; triangles, L mode; open symbols, 1-ms pulse, filled symbols, 100-ms pulse.

simulations predicted that repetitive stimulation would potentiate the NR response (Fig. 4a, left panel). A similar increase in the peak current amplitude by a second stimulus was documented previously for native NRs residing in single dendritic spines² and was interpreted as non-saturation of NMDA receptors by a single pulse. The kinetics of the neurotransmitter concentration transient at central synapses is still controversial and can vary at different synapses^{4,5,7}. To verify the behaviour predicted by our model, we recorded currents elicited by fast (1 ms) applications of saturating glutamate (1 mM; see Fig. 3b, c) from receptors residing in outside-out patches. A double-pulse protocol consistently resulted in higher currents from the same patch than single-pulse stimulation (Fig. 4a, right panel). Our results therefore show that even for saturating (1 mM) neurotransmitter pulses of relatively long duration (1 ms), repetitive stimulation will elicit increased responses from NRs. Our experiments do not directly address the issue of receptor saturation at synapses but they are highly relevant to this debate.

To explore the mechanism underlying this previously unsuspected behaviour we next examined the dynamics with which the receptor occupies various kinetic states after stimulation (Fig. 4b). Glutamate binding is fast ($\sim 50 \mu\text{s}$ at 1 mM) and binding sites for the transmitter are occupied rapidly after a single pulse (half-occupancy of C_1 occurs at $100 \mu\text{s}$). At the end of the 1-ms pulse, almost all receptors still reside in a closed state (C_1 , mean lifetime $\sim 4 \text{ ms}$)

from which the neurotransmitter can still dissociate. Progress towards opening is relatively slow (127 s^{-1}) and is about the same as the rate of glutamate release back into the synaptic cleft (120 s^{-1}). Thus, even after a pulse that fully saturates the binding sites for neurotransmitter, only $\sim 50\%$ of the initially available receptors proceed to generate current, whereas the rest repopulate states that can again bind transmitter and respond to subsequent synaptic pulses. The time course for receptor return into mono-liganded and un-liganded receptor states (C^M , C^U), which represent receptor species that can be engaged by subsequent stimulation, is illustrated in Fig. 4b.

Because the number of receptors stimulated by each short pulse of neurotransmitter depends on the number of receptors with sites available for binding transmitter molecules in that pulse, the increase in current after subsequent pulses will depend on the stimulation frequency. Simulated responses to trains of stimuli show that maximal peak NR currents will be achieved only with multiple consecutive high-frequency (at least 50 Hz) 1-ms pulses (Fig. 4c). Furthermore, switching the stimulation frequency from low (10 Hz) to high (100 Hz) markedly changes the shape of the postsynaptic response from pulsatile to sustained, and the peak amplitude approximately doubles.

Both the amplitude of the NR response to a single synaptic pulse and the ability of NRs to discriminate stimulation frequency depend on the ratio between the rates for agonist dissociation and for the conformational change towards higher-affinity states. This ratio is subject to natural selection and perhaps to environmental control. The kinetic properties of NRs vary with subunit composition, neuronal preparation, location (synaptic versus extrasynaptic) and with environmental factors such as pH, bivalent cations, temperature and protein partners²³. These differences might be reflected in characteristic gating/dissociation ratios, allowing NRs to discriminate distinct frequency ranges. The NR activation reaction might provide a mechanism that links stimulation frequency to the pattern of postsynaptic calcium transients that determine either synaptic depression or potentiation²⁴. Thus, the molecular machinery of neurotransmitter receptors determines not only the shape of the postsynaptic response to a single stimulus but also the receptors' response to trains of impulses, in a frequency-dependent manner. □

Methods

Receptor expression

HEK293 cells were co-transfected with plasmids encoding rat NR1-1a, rat NR2A and green fluorescent protein (1:1:2), as described previously¹⁵. After transfection for 2 h, cells were washed and maintained in fresh medium containing 0.5 mM DL-2-amino-5-phosphopentanoic acid (AP5) and 2 mM MgCl_2 . Currents were recorded 24–48 h thereafter.

Electrophysiology

Single-channel currents were recorded at 23°C in the cell-attached configuration, as described previously¹⁵. The pipette solution contained (in mM): 150 NaCl, 2.5 KCl, 10 N-(2-hydroxyethyl) piperazine-N'-(4-butanedisulfonic acid) (HEPES), 1 EDTA, 0.1 glycine and glutamate as indicated, and was adjusted to pH 8 with NaOH. Transfected cells were identified by fluorescence microscopy and cells with the lowest non-zero level of expression were selected for patching, to increase the chances of isolating only one active receptor in the patch.

Recordings from excised outside-out patches were made with an EPC9 amplifier (HEKA Electronics Inc., Southboro, Massachusetts, USA) as described previously²⁵. The cells were constantly superfused with normal external solution at a rate of 1 ml min^{-1} . The external medium was (in mM): 150 NaCl, 2.5 KCl, 0.5 AP5, 0.1 glycine, 5 glucose and 10 HEPES (pH adjusted to 8.0 with NaOH). Pipettes were filled with a solution containing (in mM): 120 CsF, 33 KOH, 2 MgCl_2 , 1 CaCl_2 , 0.1 spermine, 10 HEPES and 11 EGTA (pH adjusted to 7.4 with CsOH). Glutamate was applied in external solution with theta pipettes mounted on a piezoelectric bimorph as described previously²⁵. Patches were positioned near the interface of the solutions flowing from adjacent barrels, and the interface was moved by applying voltage across the bimorph. The rate of solution exchange estimated from open-tip potentials was $100\text{--}200 \mu\text{s}$. Glutamate-evoked currents were analogue low-pass filtered at 3 kHz (four-pole Bessel type, -3 dB), sampled at 50 kHz and written directly to the hard drive of the computer.

Single-channel data processing

Single-channel currents were amplified (PC-505; Warner Instrument Corporation,

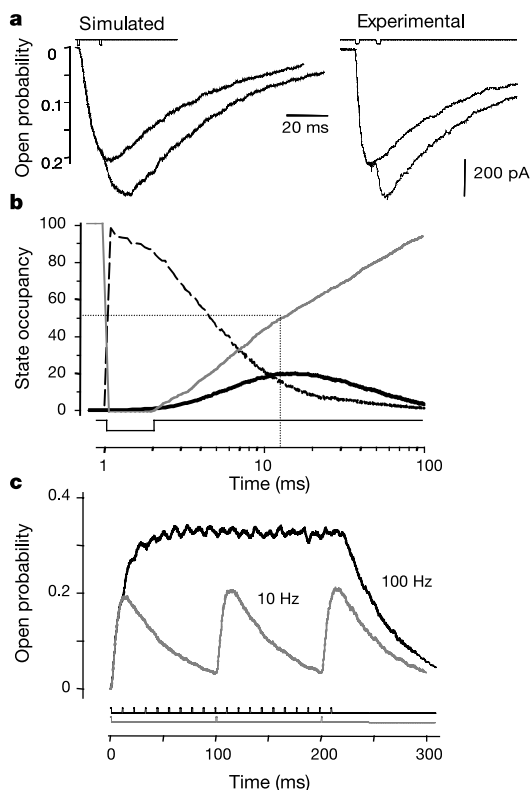


Figure 4 NR response to repetitive stimulation (1-ms pulses of 1 mM glutamate).

a, Response to one pulse is submaximal and a second, closely spaced (10-ms interval) pulse elicits a larger current. Left, responses simulated with L-mode rates and the model in Fig. 2c. Right, currents recorded from an outside-out patch. The ratio of the peak currents was 1.30 ± 0.07 (s.e.m.), $n = 6$ patches. The top traces illustrate the open-tip response. **b**, Simulated population dynamics of receptor distribution between kinetic states after one pulse. All receptors rapidly occupy a non-conducting diliganded state (C_1 , dashed line); $\sim 50\%$ release the agonist before opening once and repopulate states that can be activated by subsequent stimuli (C^M and C^U , grey line). All receptors eventually repopulate C^M and C^U ; only about half of these contributed to the peak current, which occurred at $\sim 13 \text{ ms}$ (solid black line). **c**, NRs generate currents with distinct amplitudes and temporal patterns depending on stimulation frequency.

Hamden, Connecticut, USA), low-pass filtered at 20 kHz (four-pole Bessel) and digitized at 40 kHz. The pipette potential was +100 mV (estimated membrane potential -100 to -130 mV). The currents were acquired, analysed and simulated with QUB software (www.qub.buffalo.edu). Only records from patches containing a single active channel (no double openings) were selected for processing and analysis. Idealization was performed with the segmentation *k*-means hidden Markov algorithm²⁶ after digital low-pass filtering to 12 kHz. Kinetic modelling of the idealized intervals was performed with maximum-interval LL method²⁷ using an imposed dead time of 75 μ s.

Single-channel kinetic analysis

Periods when the channel was actively gating with homogeneous kinetics were selected from each record with the following procedure. First, all intervals were fitted by kinetic schemes of increasing numbers of closed and open states to determine the minimum number of each (the threshold was >10 LL units per added state). A critical time, t_{crit} , was calculated on the basis of the time constant and amplitudes of the two slowest closed-interval distribution components²⁸. Closed intervals longer than t_{crit} were removed and the remaining intervals were joined to create an 'active-time' record. This was then divided into 1-s segments and MOTs were estimated for each segment. Segments were sorted according to their MOT and those with a MOT within one s.d. of the mean for their group were classified as H, M or L currents and analysed further.

Evaluation of kinetic models for best fit was performed with a maximum-likelihood criterion for the M and L currents²⁹. Currents obtained at the concentrations indicated in the text for each class of segments were fitted together by kinetic schemes expanded to include glutamate binding steps (see Supplementary Table S1 and Fig. S1). These schemes consisted of a core 3C2O model to which two sequential agonist-binding reactions were appended with the following constraints: $2k_{CU \rightarrow CM} = k_{CM \rightarrow C}$ and $2k_{C \rightarrow CM} = k_{CM \rightarrow CU}$. All other rates were allowed to vary freely during optimization.

Simulations

Responses were simulated from 1,000 channels by using a sampling interval of 10 kHz with the kinetic schemes shown in Fig. 2c. All channels occupied the closed unliganded state C^U before pulse application. To simulate transmitter pulses, the glutamate concentration was stepped from 0 to 1 mM for the indicated duration. Dose-response profiles were constructed from the peak open amplitude of responses elicited by square pulses with glutamate concentrations and durations as indicated. EC_{50} values were calculated by fitting normalized peak responses by the logistic equation.

Received 17 April; accepted 22 June 2004; doi:10.1038/nature02775.

- Patneau, D. K. & Mayer, M. L. Structure-activity relationships for amino acid transmitter candidates acting at *N*-methyl-D-aspartate and quisqualate receptors. *J. Neurosci.* **10**, 2385–2399 (1990).
- Mainen, Z. F., Malinow, R. & Svoboda, K. Synaptic calcium transients in single spines indicate that NMDA receptors are not saturated. *Nature* **399**, 151–155 (1999).
- Anson, L. C., Chen, P. E., Wyllie, D. J., Colquhoun, D. & Schoepfer, R. Identification of amino acid residues of the NR2A subunit that control glutamate potency in recombinant NR1/NR2A NMDA receptors. *J. Neurosci.* **18**, 581–589 (1998).
- Clements, J. D. Transmitter timecourse in the synaptic cleft: its role in central synaptic function. *Trends Neurosci.* **19**, 163–171 (1996).
- Diamond, J. S. & Jahr, C. E. Transporters buffer synaptically released glutamate on a submillisecond time scale. *J. Neurosci.* **17**, 4672–4687 (1997).
- Frerking, M. & Wilson, M. Saturation of postsynaptic receptors at central synapses? *Curr. Opin. Neurobiol.* **6**, 395–403 (1996).
- Franks, K. M., Bartol, T. M. Jr & Sejnowski, T. J. A Monte Carlo model reveals independent signaling at central glutamatergic synapses. *Biophys. J.* **83**, 2333–2348 (2002).
- Min, M. Y., Rusakov, D. A. & Kullmann, D. M. Activation of AMPA, kainate, and metabotropic receptors at hippocampal mossy fiber synapses: role of glutamate diffusion. *Neuron* **21**, 561–570 (1998).
- McAllister, A. K. & Stevens, C. F. Nonsaturation of AMPA and NMDA receptors at hippocampal synapses. *Proc. Natl Acad. Sci. USA* **97**, 6173–6178 (2000).
- Ishikawa, T., Sahara, Y. & Takahashi, T. A single packet of transmitter does not saturate postsynaptic glutamate receptors. *Neuron* **34**, 613–621 (2002).
- Oertner, T. G., Sabatini, B. L., Nimchinsky, E. A. & Svoboda, K. Facilitation at single synapses probed with optical quantal analysis. *Nature Neurosci.* **5**, 657–664 (2002).
- Nowak, L., Bregestovski, P., Ascher, P., Herbet, A. & Prochiantz, A. Magnesium gates glutamate-activated channels in mouse central neurones. *Nature* **307**, 462–465 (1984).
- Traynelis, S. F. & Cull-Candy, S. G. Proton inhibition of *N*-methyl-D-aspartate receptors in cerebellar neurones. *Nature* **345**, 347–350 (1990).
- Westbrook, G. L. & Mayer, M. L. Micromolar concentrations of Zn^{2+} antagonize NMDA and GABA responses of hippocampal neurones. *Nature* **328**, 640–643 (1987).
- Popescu, G. & Auerbach, A. Modal gating of NMDA receptors and the shape of their synaptic response. *Nature Neurosci.* **6**, 476–483 (2003).
- Magleby, K. L. Modal gating of NMDA receptors. *Trends Neurosci.* **27**, 231–233 (2004).
- Popescu, G. & Auerbach, A. The NMDA receptor gating machine: lessons from single channels. *Neuroscientist* **10**, 192–198 (2004).
- Anson, L. C., Schoepfer, R., Colquhoun, D. & Wyllie, D. J. Single-channel analysis of an NMDA receptor possessing a mutation in the region of the glutamate binding site. *J. Physiol. (Lond.)* **527**, 225–237 (2000).
- Colquhoun, D. Binding, gating, affinity and efficacy: the interpretation of structure-activity relationships for agonists and of the effects of mutating receptors. *Br. J. Pharmacol.* **125**, 924–947 (1998).
- Chen, N., Ren, J., Raymond, L. A. & Murphy, T. H. Changes in agonist concentration dependence that are a function of duration of exposure suggest *N*-methyl-D-aspartate receptor nonsaturation during synaptic stimulation. *Mol. Pharmacol.* **59**, 212–219 (2001).
- Carmignoto, G. & Vicini, S. Activity-dependent decrease in NMDA receptor responses during development of the visual cortex. *Science* **258**, 1007–1011 (1992).

- Vicini, S. *et al.* Functional and pharmacological differences between recombinant *N*-methyl-D-aspartate receptors. *J. Neurophysiol.* **79**, 555–566 (1998).
- Cull-Candy, S., Brickley, S. & Farrant, M. NMDA receptor subunits: diversity, development and disease. *Curr. Opin. Neurobiol.* **11**, 327–335 (2001).
- Yang, S. N., Tang, Y. G. & Zucker, R. S. Selective induction of LTP and LTD by postsynaptic $[Ca^{2+}]_i$ elevation. *J. Neurophysiol.* **81**, 781–787 (1999).
- Robert, A., Irizarry, S. N., Hughes, T. E. & Howe, J. R. Subunit interactions and AMPA receptor desensitization. *J. Neurosci.* **21**, 5574–5586 (2001).
- Qin, F. Restoration of single-channel currents using the segmental *k*-means method based on hidden Markov modeling. *Biophys. J.* **86**, 1488–1501 (2004).
- Qin, F., Auerbach, A. & Sachs, F. Maximum likelihood estimation of aggregated Markov processes. *Proc. R. Soc. Lond. B* **264**, 375–383 (1997).
- Magleby, K. L. & Pallotta, B. S. Burst kinetics of single calcium-activated potassium channels in cultured rat muscle. *J. Physiol. (Lond.)* **344**, 605–623 (1983).
- Qin, F., Auerbach, A. & Sachs, F. A direct optimization approach to hidden Markov modeling for single channel kinetics. *Biophys. J.* **79**, 1915–1927 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank T. Bailey, M. Teeling and C. Nicolai for technical assistance. This work was supported by NIH grants to G.P. and A.A.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.P. (popescu@buffalo.edu).

Regulation of innate and adaptive immune responses by MAP kinase phosphatase 5

Yongliang Zhang¹†, Joseph N. Blattman¹, Norman J. Kennedy², Julie Duong¹, Thang Nguyen¹, Ying Wang¹, Roger J. Davis², Philip D. Greenberg¹, Richard A. Flavell³* & Chen Dong¹*†

¹Department of Immunology, University of Washington, Seattle, Washington 98195-7650, USA

²Howard Hughes Medical Institute, University of Massachusetts, Worcester, Massachusetts 01605, USA

³Section of Immunobiology, Howard Hughes Medical Institute, Yale University, New Haven, Connecticut 06520, USA

* These authors contributed equally to this work

† Present address: Department of Immunology, University of Texas, MD Anderson Cancer Center, Houston, Texas 77030, USA

Mitogen-activated protein (MAP) kinases are essential regulators in immune responses¹, and their activities are modulated by kinases and phosphatases. MAP kinase phosphatase (MKP) is a family of dual-specificity phosphatases whose function is evolutionarily conserved^{2,3}. A number of mammalian MKPs have been identified so far^{2,3}, but their specific physiological functions in negative regulation of MAP kinases have not been genetically defined. Here we examine innate and adaptive immune responses in the absence of MKP5. JNK activity was selectively increased in *Mkp5* (also known as *Dusp10*)-deficient mouse cells. *Mkp5*-deficient cells produced greatly enhanced levels of pro-inflammatory cytokines during innate immune responses and exhibited greater T-cell activation than their wild-type counterparts. However, *Mkp5*-deficient T cells proliferated poorly upon activation, which resulted in increased resistance to experimental autoimmune encephalomyelitis. By contrast, *Mkp5*-deficient CD4⁺ and CD8⁺ effector T cells produced significantly increased levels of cytokines compared with wild-type cells, which led to much more robust and rapidly fatal immune responses to secondary infection with lymphocytic choriomeningitis virus. Therefore, MKP5 has a principal function in both innate and adaptive

immune responses, and represents a novel target for therapeutic intervention of immune diseases.

To examine MKP function in immune responses, we first sought to identify the mammalian homologue of the fly *puckered* gene that encodes a JNK-specific dual-specificity phosphatase⁴, and found that human MKP5, previously shown to inactivate p38 and JNK *in vitro*^{5,6}, exhibits the most significant sequence homology (data not shown). We subsequently cloned mouse *Mkp5* complementary DNA based on the sequences detected in the expressed sequence tag database (data not shown). As a first step to analyse MKP5 function in immune responses, its expression was examined in immune cells using polymerase chain reaction with reverse transcription (RT-PCR). The mouse macrophage cell line RAW does not constitutively express MKP5, but expression was strongly induced as early as 15 min after lipopolysaccharide (LPS) treatment (Supplementary Fig. 1a). In the T-cell compartment, MKP5 was constitutively expressed by naive CD4⁺ T cells; however, it was downregulated 24 h after T-cell activation (Supplementary Fig. 1a). MKP5 expression was selectively higher in CD4⁺ Th2 than Th1 cells (Supplementary Fig. 1a), correlating with stronger JNK and p38 activities in Th1 cells^{7,8}.

As the AP1 transcription factor is a major target of MAP kinase pathways, we tested whether MKP5 inhibited AP1-dependent transcription activity. A full-length mouse MKP5 cDNA was expressed in RAW cells, which led to inhibition of constitutive and LPS-induced AP1 reporter activities (Supplementary Fig. 1b). Furthermore, the activity of the tumour-necrosis factor (TNF)- α promoter, which is regulated by AP1 factors, was also strongly inhibited in RAW cells by MKP5. In human Jurkat cells, MKP5 reduced AP1 activity, and accordingly AP1-regulated interferon (IFN)- γ and interleukin (IL)-4 promoter activities (Supplementary Fig. 1c). By contrast, a phosphatase-dead mutant of MKP5, expressed at equivalent levels as the wild-type molecule, did not inhibit AP1 activity (Supplementary Fig. 1b, c). Therefore, MKP5

can negatively regulate transcriptional activity mediated by AP1 factors in immune cells.

To understand the role of MKP5 in regulating MAP kinase during the immune response, we created *Mkp5*-deficient (*Mkp5*^{-/-}) mice through homologous recombination in embryonic stem (ES) cells (Fig. 1a), and homozygous *Mkp5* knockout mice were subsequently obtained. Messenger RNA analysis of *Mkp5* expression in T cells derived from *Mkp5*^{+/-} and *Mkp5*^{-/-} mice confirmed that an *Mkp5*-null mutation had been obtained (Fig. 1b), and indicated that unlike the fly *puckered* gene, *Mkp5* is not essential for development. Lymphoid and myeloid cells developed normally in the absence of the *Mkp5* gene (data not shown). Therefore, MKP5 is not required for development of the immune system.

As MKP5 has been shown to be a negative regulator of p38 and JNK *in vitro*^{5,6}, we investigated whether MKP5 is required to inactivate JNK, p38 or both in effector T cells using *in vitro* kinase assays, as described previously^{9,10}. Both Th1 and Th2 cells deficient in MKP5 exhibited greatly enhanced levels of JNK activity but no enhancement of p38 or NF- κ B activity (Fig. 1c and data not shown), indicating that MKP5 negatively regulates JNK *in vivo*. However, this enhanced JNK activity in *Mkp5*^{-/-} T cells was rapidly down-regulated after re-stimulation with anti-CD3, demonstrating the presence of an MKP5-independent mechanism that can terminate the JNK pathway in effector T cells. These results were further confirmed by western blotting with anti-phospho-JNK and p38 antibodies (data not shown). Increased JNK activity was also observed in *Mkp5*^{-/-} macrophages after LPS treatment (data not shown). Therefore, MKP5 was required to limit JNK activity in innate and adaptive immune responses.

We then examined immune responses in *Mkp5*-deficient animals. As the *Mkp5* gene is rapidly upregulated in macrophages after LPS treatment, MKP5 function in response to LPS stimulation was studied. Peritoneal macrophages from *Mkp5*^{+/+} and *Mkp5*^{-/-} mice were stimulated with different doses of LPS. Compared with wild-type macrophages, *Mkp5*^{-/-} macrophages expressed increased mRNA and protein levels of the pro-inflammatory cytokines IL-6 and TNF- α (Fig. 2a, b). *Mkp5*^{-/-} cells exhibited enhanced cytokine expression at all time points examined, although the overall kinetics remained the same (Fig. 2c). Injection of LPS into *Mkp5*^{-/-} mice also led to increased serum TNF- α production compared with wild-type controls (Fig. 2d).

LPS induces innate responses through Toll-like receptor (TLR) 4 (ref. 11). To examine the role of MKP5 in regulation of other TLR signals, we stimulated wild-type and *Mkp5*^{-/-} macrophages with peptidoglycan (PGN) and poly(I:C), which trigger TLR2 (ref. 12) and TLR3 (ref. 13), respectively. *Mkp5*^{-/-} macrophages produced increased levels of IL-6 and TNF- α in response to these two pathogen-related molecules (Supplementary Fig. 2a, b). Furthermore, after infection with *Listeria monocytogenes*, a Gram-positive bacterium, *Mkp5*^{-/-} macrophages produced significantly greater levels of IL-6 and TNF- α compared with wild-type macrophages (Supplementary Fig. 2c).

In addition to production of pro-inflammatory cytokines, another important function of the innate immune system is activation of antigen-specific T cells¹¹. We thus examined the T-cell priming capacities of *Mkp5*^{-/-} antigen-presenting cells (APCs). Wild-type and knockout APCs were similarly inefficient in priming T cells in the absence of innate activation (data not shown). After LPS treatment, however, *Mkp5*^{-/-} APCs exhibited greatly enhanced priming of OT-I and OT-II cells with regard to both IL-2 production and the proliferative response (Fig. 2e). Together, the above data indicate that MKP5 is a negative regulator of innate immunity.

MAP kinases are also crucial regulators of Th cell activation and effector cytokine expression¹. When CD4 T cells from wild-type or *Mkp5*^{-/-} mice were activated by plate-bound anti-CD3 with or without anti-CD28 co-stimulation, *Mkp5*^{-/-} CD4 T cells consistently showed reduced proliferation after activation (Fig. 3a),

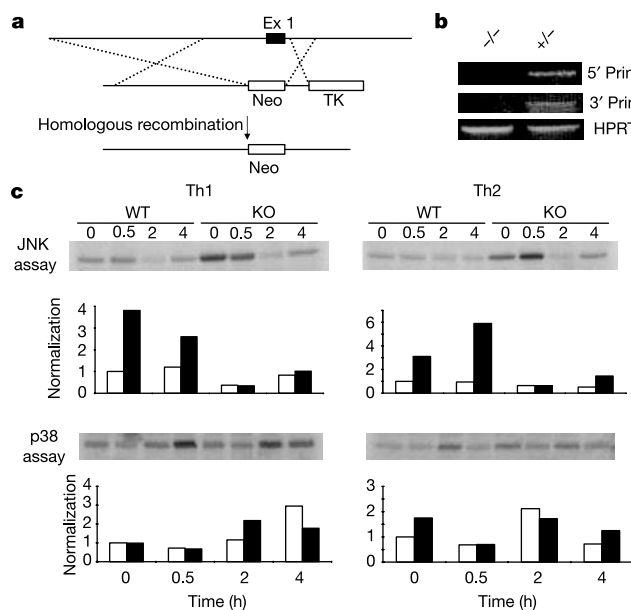


Figure 1 Targeted disruption of the mouse *Mkp5* gene results in increased JNK activity. **a**, Diagram of partial structure of the mouse *Mkp5* gene, the targeting vector and the disrupted allele. **b**, RT-PCR analysis of *Mkp5* gene expression in *Mkp5*^{+/+} and *Mkp5*^{-/-} T cells using primers specific for sequences encoded by exon 1 (5') or downstream exons (3'). **c**, Wild-type (WT; open bars) and knockout (KO; filled bars) Th1 and Th2 cells were re-stimulated with anti-CD3 for the indicated times and cell lysates were prepared to examine the JNK and p38 MAP kinase activities using glutathione S-transferase (GST)-c-Jun or GST-ATF2, respectively, as substrates.

indicating that MKP5 is required for proper T-cell expansion. This result is the reciprocal of our previous finding that T cells deficient in JNK proliferate better than wild-type cells^{9,10}.

To examine the function of MKP5 in regulating effector T-cell

function, we differentiated naive wild-type or knockout CD4 T cells into Th1 and Th2 cells. *Mkp5*^{-/-} Th1 and Th2 cells produced markedly increased levels of IFN- γ and IL-4, respectively (Fig. 3b). Intracellular staining of cytokines revealed that both the total

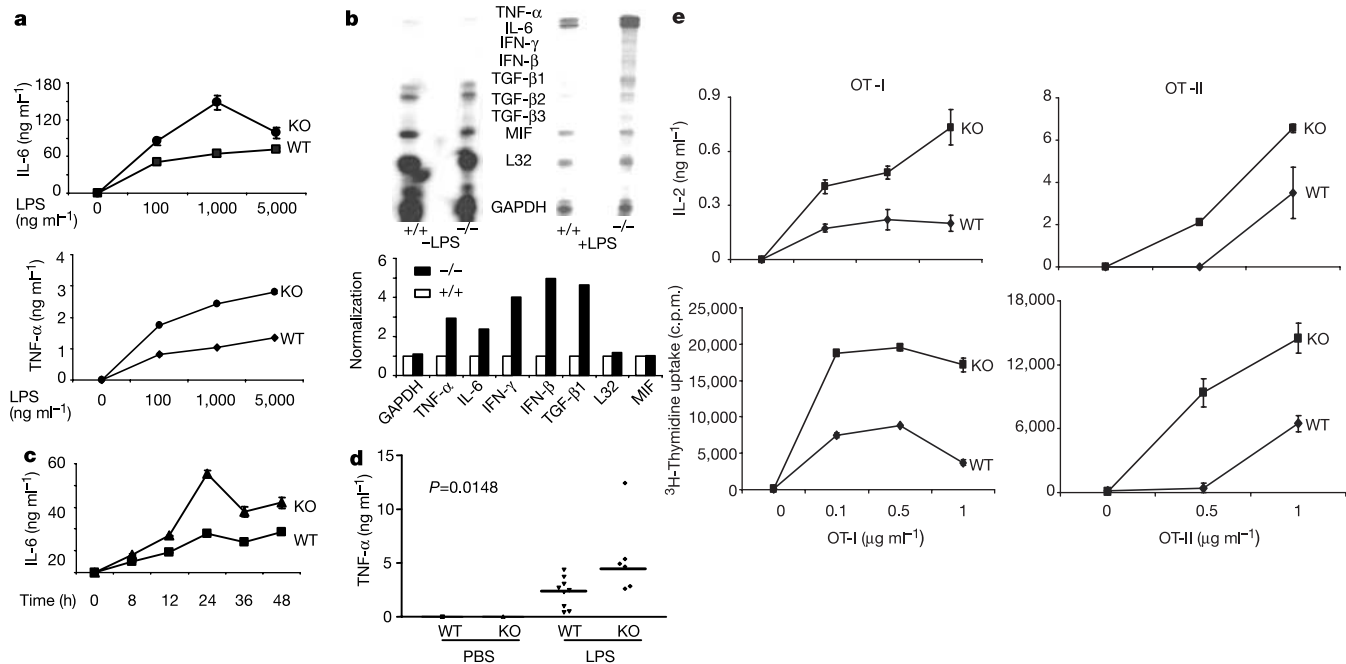


Figure 2 Increased innate immune responses in the absence of MKP5. **a**, IL-6 and TNF- α production by wild-type and *Mkp5*^{-/-} peritoneal macrophages upon LPS stimulation. **b**, Cytokine mRNA was measured in LPS-treated wild-type and *Mkp5*^{-/-} macrophages after 8 h treatment. MIF, macrophage migration inhibitory factor. **c**, IL-6 production at varying time points by wild-type and *Mkp5*^{-/-} peritoneal macrophages after incubation

with 100 ng ml⁻¹ LPS. **d**, TNF- α production in wild-type and *Mkp5*^{-/-} mice serum after 1.5 h LPS injection. **e**, Priming of OT-I and OT-II cells by wild-type and *Mkp5*^{-/-} APCs was determined by measuring IL-2 production and proliferation. Data are from one of three independent experiments with similar results. Error bars represent standard deviations of triplicate samples. c.p.m., counts per minute.

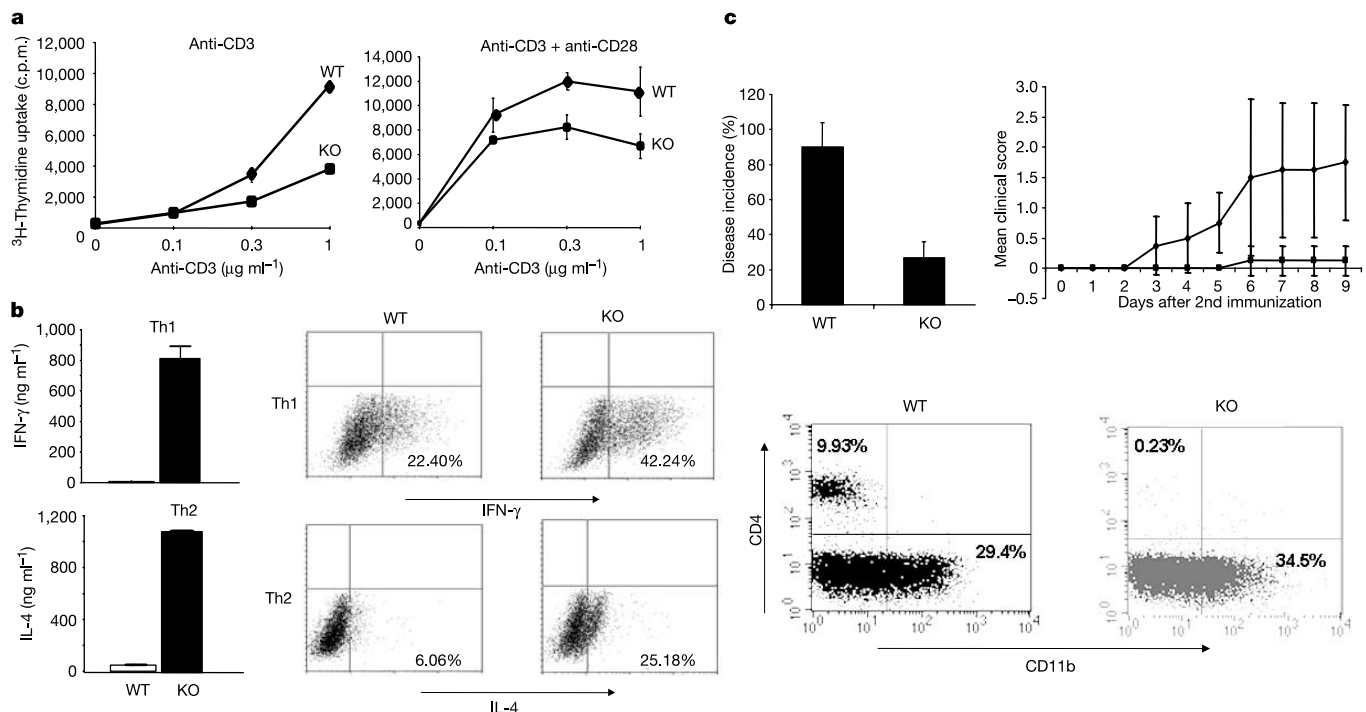


Figure 3 Defective CD4 T-cell responses by *Mkp5*^{-/-} mice. **a**, Naive CD4 T cells isolated from wild-type and knockout mice were stimulated with anti-CD3 with or without anti-CD28 for three days. ³H-thymidine was added into the culture for the last 8 h of incubation. **b**, Cytokine production after stimulation of wild-type and knockout Th1 and Th2 cells measured

by ELISA and intracellular staining. **c**, EAE was induced in wild-type and *Mkp5*^{-/-} mice. The percentage of mice developing disease was based on two independent experiments. Average disease scores from diseased mice are shown from a representative of two experiments. Error bars represent standard deviations of triplicate samples.

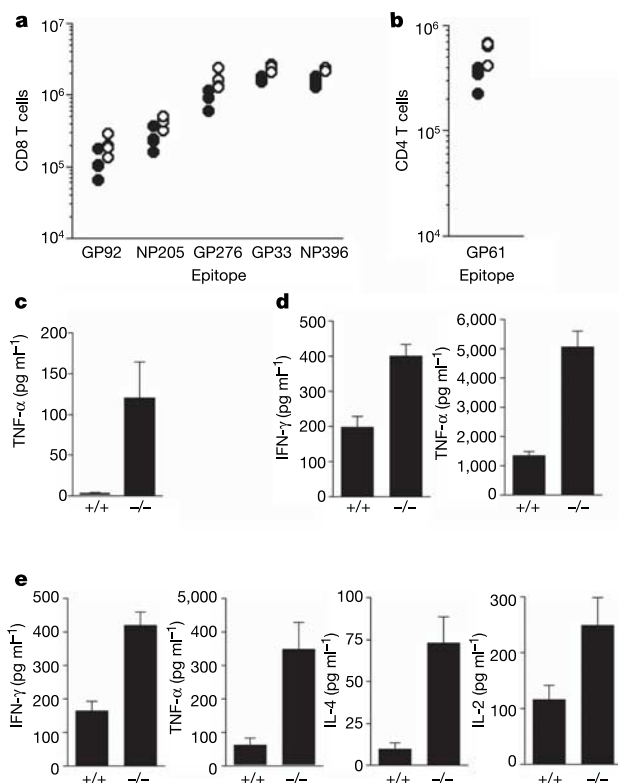


Figure 4 Enhanced T-cell responses in *Mkp5*^{-/-} mice after secondary LCMV infection. **a, b**, Wild-type (open circles) or *Mkp5*^{-/-} (filled circles) mice were infected with LCMV Armstrong and re-challenged 28 days later with LCMV clone 13. Virus-specific CD8 (**a**) and CD4 (**b**) T cells were quantified 4 days after re-challenge by intracellular cytokine staining. **c**, Serum TNF-α levels measured in LCMV-infected mice 4 days after re-challenge. **d, e**, Cytokine production by CD8 (**d**) and CD4 (**e**) splenocytes from wild-type or *Mkp5*^{-/-} mice stimulated with defined peptide epitopes as determined by cytokine bead array. Error bars represent standard deviations of multiple samples.

number of cytokine-producing cells and cytokine expression per cell were greatly increased in the absence of MKP5 (Fig. 3b). We also found that *Mkp5*^{-/-} effector CD8 T cells produced more IFN-γ and TNF-α *in vitro* (data not shown). These results indicate that MKP5 is a negative regulator of effector T-cell cytokine expression.

To substantiate the above findings, we characterized *Mkp5*-knockout mice in several *in vivo* immune and autoimmune models. Wild-type and *Mkp5*^{-/-} mice were immunized with keyhole limpet haemocyanin (KLH) in complete Freund's adjuvant (CFA), and T-cell proliferation and cytokine production was determined after KLH re-stimulation *ex vivo*. Although antigen-driven proliferation of splenic T cells from immunized *Mkp5*^{-/-} mice was decreased, *Mkp5*^{-/-} cells produced increased levels of IFN-γ and IL-4 compared with the wild-type cells (Supplementary Fig. 3a, b). This again demonstrates the differential roles of MKP5 in regulation of T-cell clonal expansion and effector T-cell cytokine expression.

To examine the effect of MKP5 deficiency in T-cell-mediated immune disease, we immunized wild-type and *Mkp5*^{-/-} mice with myelin oligodendrocyte glycoprotein (MOG) peptide to induce experimental autoimmune encephalomyelitis (EAE). Compared with wild-type mice, *Mkp5*^{-/-} mice exhibited reduced incidence and severity of disease (Fig. 3c). Consistently, the number of CD4 T cells in the brain of *Mkp5*^{-/-} mice was greatly reduced (Fig. 3c). This experiment suggests a crucial role of MKP5 in the generation and/or expansion of autoreactive T cells in EAE disease.

To examine the role of MKP5 in T-cell-mediated immunity to infection, we infected wild-type and *Mkp5*^{-/-} mice with LCMV. *Mkp5*^{-/-} mice cleared virus and showed little difference in the

primary T-cell response to LCMV (data not shown), suggesting that MKP5 is dispensable for T-cell activation in the context of a live viral infection. Upon secondary LCMV challenge, the *Mkp5*^{-/-} mice exhibited a minor increase in the number of responding virus-specific CD8 and CD4 T cells (Fig. 4a, b). However, as previously observed, these responding T cells produced markedly more effector cytokines (Fig. 4d, e). The increased production of growth factors such as IL-2 and IL-4 may have accounted for the slight increase in the number of responding T cells. The most notable phenotype reflected the increased production of TNF-α by both CD4 and CD8 cells, which yielded markedly elevated levels of serum TNF-α in *Mkp5*^{-/-} mice compared with wild-type mice, and was probably responsible for the observed immune-mediated death of these mice at 2–4 days after re-challenge (Fig. 4c and data not shown).

Our data indicate that MKP5 negatively regulates the JNK signalling pathway and has an important role in immune responses. It is induced during innate immune responses and serves as a negative regulator. MKP5 is also required for proper T-cell proliferation, and its deficiency affords some protection from the development of autoimmune EAE disease. Additionally, MKP5 negatively regulates the function of effector CD4 and CD8 T cells, which may be essential for protecting the host from injury by excessive T-cell responses to pathogens. This work provides definitive evidence by means of a genetic approach to support a non-redundant role for a mammalian dual-specificity phosphatase in negative regulation of MAP kinases. This analysis not only demonstrates the importance of MAP kinase phosphatases in innate and adaptive immune responses, but also further reveals the complexity of signal transduction mechanisms in immune responses. □

Methods

Generation of *Mkp5*^{-/-} mice

An *Mkp5* gene-targeting construct that replaces exon 1 (encoding the first approximately 250 amino acids) and its 5' flanking sequences with a neomycin-resistance gene was transfected into TC-1 ES cells. Two homologous recombinants were used to inject C57BL/6 blastocysts. Male chimaera mice were used to generate homozygous knockout animals.

Innate immunity assays

Peritoneal macrophages were isolated and treated with LPS from *Escherichia coli* (Sigma). IL-6 and TNF-α in the supernatant were measured at 24 h by enzyme-linked immunosorbent assay (ELISA), and cytokine mRNA at 8 h by RNase protection assay (BD Pharmingen). For LPS injection, wild-type and *Mkp5*^{-/-} mice were injected intraperitoneally with PBS or 300 μg *E. coli* O55:B5 LPS (normalized by body weight). Ninety minutes later, the mice were killed and serum TNF-α levels were determined by ELISA (Pharmingen). A two-tailed Mann-Whitney T-test was performed using GraphPad Prism program (v2.01) to determine the *P* value. For T-cell priming assays, wild-type and *Mkp5*^{-/-} splenic APCs were incubated with 800 ng ml⁻¹ LPS overnight, irradiated and incubated with OT-I and OT-II cells in the presence of stimulating peptides. IL-2 in the supernatant was determined at 24 h by ELISA (Pharmingen) and cell proliferation by ³H-thymidine incorporation at 72 h.

Induction of EAE

EAE was induced with MOG33–55 peptide in CFA plus pertussis toxin. Mice were immunized subcutaneously at the dorsal flanks with 100 μg of MOG peptide in CFA at day 0 and day 8. Pertussis toxin was given intraperitoneally at day 1 and day 9 with the dosage of 200 μg per mouse. Clinical scores ranged from tail weakness, grade 1; hindlimb weakness, grade 2; through to partial hindlimb paralysis, grade 3. To analyse central nervous system infiltrates, both brain and spinal cord were collected and mononuclear cells prepared by percoll gradient. All cells were stained by anti-CD4 and anti-Mac-1, and flow cytometry was used to analyse central nervous system infiltrating mononuclear cells in wild-type and knockout mice.

LCMV infection and immune analysis

Mice were inoculated with 2×10^5 plaque-forming units (PFU) of the Armstrong strain of LCMV, and re-challenged 28 days later with 2×10^6 PFU of the clone-13 strain of LCMV. Intracellular cytokine staining for IFN-γ was performed after 5 h stimulation with the indicated peptide. Cells were cultured with either known CD8 epitopes (GP276, GP33 and NP396, all 0.1 μM) or a CD4 epitope (GP61, 1 μM) for 36 h and cytokine production was measured by bead array analysis (Beckton Dickinson).

Received 5 April; accepted 17 June 2004; doi:10.1038/nature02764.

1. Dong, C., Davis, R. J. & Flavell, R. A. MAP kinases in the immune response. *Annu. Rev. Immunol.* **20**, 55–72 (2002).

2. Camps, M., Nichols, A. & Arkinstall, S. Dual specificity phosphatases: a gene family for control of MAP kinase function. *FASEB J.* **14**, 6–16 (2000).
3. Theodosiou, A. & Ashworth, A. MAP kinase phosphatases. *Genome Biol.* **3**, 3009–3018 (2002).
4. Martin-Blanco, E. *et al.* pucker encodes a phosphatase that mediates a feedback loop regulating JNK activity during dorsal closure in *Drosophila*. *Genes Dev.* **12**, 557–570 (1998).
5. Theodosiou, A., Smith, A., Gillieron, C., Arkinstall, S. & Ashworth, A. MKP5, a new member of the MAP kinase phosphatase family, which selectively dephosphorylates stress-activated kinases. *Oncogene* **18**, 6981–6988 (1999).
6. Tanoue, T., Moriguchi, T. & Nishida, E. Molecular cloning and characterization of a novel dual specificity phosphatase, MKP-5. *J. Biol. Chem.* **274**, 19949–19956 (1999).
7. Rincon, M., Derijard, B., Chow, C. W., Davis, R. J. & Flavell, R. A. Reprogramming the signalling requirement for AP-1 (activator protein-1) activation during differentiation of precursor CD4⁺ T-cells into effector Th1 and Th2 cells. *Genes Funct.* **1**, 51–68 (1997).
8. Rincon, M. *et al.* Interferon- γ expression by Th1 effector T cells mediated by the p38 MAP kinase signaling pathway. *EMBO J.* **17**, 2817–2829 (1998).
9. Dong, C. *et al.* Defective T cell differentiation in the absence of Jnk1. *Science* **282**, 2092–2095 (1998).
10. Dong, C. *et al.* JNK is required for effector T-cell function but not for T-cell activation. *Nature* **405**, 91–94 (2000).
11. Janeway, C. A. Jr & Medzhitov, R. Innate immune recognition. *Annu. Rev. Immunol.* **20**, 197–216 (2002).
12. Takeuchi, O. *et al.* Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components. *Immunity* **11**, 443–451 (1999).
13. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF- κ B by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank L. Evangelisti, C. Hughes and J. Stein for technical assistance; P. Leder for providing TC-1 ES cells; R. Alanis for his help with the *Listeria* infection experiment; A. Farr for his guidance in animal studies; and the entire Dong laboratory for their support and help. R.J.D. and R.A.F. are HHMI investigators, and C.D. is an Arthritis Investigator of the Arthritis Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A.F. (richard.flavell@yale.edu) or C.D. (cdong@mdanderson.org).

Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control

Eva Hernando^{1*}, Zaher Nahle^{3*†}, Gloria Juan¹, Elena Diaz-Rodriguez², Miguel Alaminos^{1†}, Michael Hemann³, Loren Michel², Vivek Mittal³, William Gerald¹, Robert Benezra², Scott W. Lowe³ & Carlos Cordon-Cardo¹

¹Department of Pathology and ²Cancer Biology and Genetics Program, Memorial Sloan-Kettering Cancer Center New York, New York 10021, USA

³Cold Spring Harbor Laboratory 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA

* These authors contributed equally to this work.

† Present addresses: Department of Internal Medicine, Washington University in St. Louis, Saint Louis, Missouri, Missouri 63110, USA (Z.N.); Department of Histology, School of Medicine, University of Granada, Avenida de Madrid, 11. E-18012 Granada, Spain (M.A.)

Advanced human cancers are invariably aneuploid, in that they harbour cells with abnormal chromosome numbers^{1,2}. However, the molecular defects underlying this trait, and whether they are a cause or a consequence of the malignant phenotype, are not clear. Mutations that disable the retinoblastoma (Rb) pathway are also common in human cancers¹. These mutations promote tumour development by deregulating the E2F family of transcription factors leading to uncontrolled cell cycle progression³. We show that the mitotic checkpoint protein Mad2 is a direct E2F target and, as a consequence, is aberrantly expressed in cells with

Rb pathway defects. Concordantly, Mad2 is overexpressed in several tumour types, where it correlates with high E2F activity and poor patient prognosis. Generation of Rb pathway lesions in normal and transformed cells produces aberrant Mad2 expression and mitotic defects leading to aneuploidy, such that elevated Mad2 contributes directly to these defects. These results demonstrate how chromosome instability can arise as a by-product of defects in cell cycle control that compromise the accuracy of mitosis, and suggest a new model to explain the frequent appearance of aneuploidy in human cancer.

The appearance of aneuploidy in human cancers has been linked to defects in several processes^{1,2}. Many studies suggest that aneuploidy arises from defects in the conserved spindle checkpoint that normally governs progression through mitosis^{4,5}. In response to lack of tension or improper microtubule attachment at the kinetochores, a group of sensor proteins (Bub3, Bub1, BubR1, Mps1 and Mad2) releases a diffusible signal that inhibits the ubiquitin ligase activity of the anaphase promoting complex (APC) or cyclosome⁶. APC/cyclosome function is required for sister chromatid separation and cytokinesis⁷. Although mitotic checkpoint defects are often observed in cancer cells challenged with microtubule poisons, and inactivation of checkpoint components produces aneuploidy in yeast and mammalian cells^{8–10}, loss-of-function mutations in spindle checkpoint genes are rarely observed in human tumours^{11,12}.

Adenovirus E1A is a potent viral oncoprotein that acts, in part, by inactivating the Rb gene product and deregulating the E2F transcription factors³. In performing a series of microarray experiments, we observed that E1A significantly increased the expression of *MAD2* (Z.N., V.M. and S.W.L., unpublished observations), a crucial component of the spindle checkpoint that associates with the APC/cyclosome, and prevents its activation^{7,13}. Northern and western blot analyses of *MAD2* transcript and protein confirmed its upregulation by E1A in mouse embryo fibroblasts (MEFs) (Fig. 1a, compare lanes 1 and 2 or 3 and 4, respectively). Similarly, MEFs isolated from *Rb*^{-/-} mice displayed increased Mad2 levels compared to *Rb*^{+/+} controls (Fig. 1a, compare lanes 5 and 6). IMR90 human fibroblasts expressing either *E1A* (Fig. 1b, compare lanes 1 and 2) or *E2F-1* (Fig. 1b, compare lanes 1 and 3) also expressed elevated Mad2 levels, implying that deregulation of the Rb/E2F pathway induces Mad2 through a conserved mechanism.

The above results are consistent with the possibility that *MAD2* is an E2F target. Concordantly, previous global chromatin immunoprecipitation/microarray studies suggested that E2Fs can bind the *MAD2* promoter in normal fibroblasts¹⁴. Indeed, analysis of the *MAD2* genomic sequence showed several putative E2F-binding sites in the *MAD2* promoter (see Supplementary Fig. S1), and subsequent chromatin immunoprecipitation (ChIP) analysis confirmed *in vivo* binding of E2F-1 to their predicted sites in *E1A*-expressing cells (Fig. 1c, lane 2). Moreover, a *MAD2* genomic fragment containing the E2F sites (Supplementary Fig. S1) conferred E2F responsiveness to a luciferase reporter following transient transfection into IMR90 (data not shown) or U-2OS cells (Fig. 1d) in a manner comparable to the *caspase-7* promoter, an established E2F target¹⁵. Therefore, *MAD2* is a physiological transcriptional target of E2F.

Many E2F targets are cell-cycle regulated³. Therefore, we examined the cell cycle distribution of Mad2 expression in normal synchronously and asynchronously cycling cells. The level of Mad2 was undetectable in quiescent IMR90 cells (Fig. 1e, *t* = 0) and increased after cells entered S phase following serum addition, reaching maximum levels in G2/M, albeit slightly after cyclin A. The increase in Mad2 was confined largely to mitosis, as assessed by co-expression with phosphorylated histone H3 following release from a double-thymidine-induced S-phase arrest (Fig. 1f), and by laser scanning cytometry (LSC) analysis of Mad2 expression in unsynchronized NIH-3T3 fibroblasts (Fig. 1g). Presumably, Mad2 expression is regulated during the normal cell cycle as part of a

mechanism that coordinates the expression of spindle checkpoint proteins with the onset of G2/M.

E2F target genes are often constitutively expressed in cells with Rb pathway defects. Concordantly, Mad2 was expressed at aberrantly high levels throughout the cell cycle in IMR90 cells expressing E1A (Fig. 2a), which deregulates endogenous E2F, and in cells where Rb was suppressed by stable RNA interference (RNAi) (shRb¹⁶, Fig. 2a). Of note, Mad2 levels in E1A- and shRb-expressing IMR90 cells were similar to those produced by enforced expression of MAD2 (Fig. 2b, c). As in normal fibroblasts, Mad2 was confined to G2/M in the Rb-functional T24 bladder carcinoma¹⁷ and HCT116 colon carcinoma lines (Fig. 2d). However, in bladder and colon carcinoma lines with inactive Rb, such as HT1197 (ref. 17) and SW480 (ref. 18), respectively, Mad2 was expressed at high levels throughout the cell cycle (Fig. 2d). Thus, cells with Rb defects often aberrantly express Mad2.

Because the Rb pathway is frequently disrupted during tumorigenesis, we asked whether Mad2 is overexpressed in human tumour specimens. Retinoblastomas, which invariably lose Rb function and deregulate E2F, expressed high levels of Mad2 relative to normal retina (Fig. 3a, upper panels). Similarly, Mad2 was significantly increased in many bladder carcinomas when compared to normal urothelium (data not shown), particularly in the more advanced tumours, where Rb inactivation is common¹⁹ (Fig. 3a, middle panels). The elevated Mad2 levels observed in human tumours were not merely due to the increased proliferation of cancer cells, as Mad2 was not always detected in highly proliferating normal and tumour samples (see Supplementary Fig. S2).

A clear link between deregulated E2F activity and aberrant Mad2

expression was observed in neuroblastoma. Here, MAD2 messenger RNA levels correlated tightly with E2F1 mRNA expression (an E2F member but also a prototypical E2F target gene), as shown by analysis of a microarray database containing information on 106 neuroblastoma tumours and 12 neuroblastoma cell lines²⁰ (Fig. 3b; Kendall *tau* correlation coefficient: 0.568; $P < 0.0001$) and validated by immunohistochemistry (Fig. 3a, lower panels). Moreover, aberrant Mad2 expression was associated with MYCN (a proto-oncogene of the myelocytomatosis (MYC)-box genes) overexpression and amplification (which can drive E2F constitutive activation²¹) and poor patient prognosis (Fig. 3c). In fact, high Mad2 expression was more tightly associated with survival than MYCN amplification ($P = 0.001$ for MAD2; $P = 0.0215$ for MYCN). Interestingly, MAD2 also clusters with tens of genes capable of predicting disease outcome in breast cancer patients²². Therefore, aberrant Mad2 expression is linked to Rb pathway defects in human tumours, where it can be an effective indicator of patient prognosis.

Most advanced human cancers are aneuploid, and loss of Rb function has been associated with mitotic defects²³. Hence, it is paradoxical that cells with defective Rb express higher levels of a protein that should facilitate proper checkpoint responses. We hypothesized that misexpressed or elevated Mad2 might compromise mitotic events, predisposing cells to genomic instability. Therefore, we examined the DNA content of IMR90 fibroblasts expressing E1A, shRb or Mad2, all of which expressed similarly high levels of Mad2 throughout the cell cycle (see Fig. 2). Biparametric LSC analysis comparing BrdU incorporation and DNA content demonstrated that E1A-, shRB- and MAD2-transduced IMR90 or

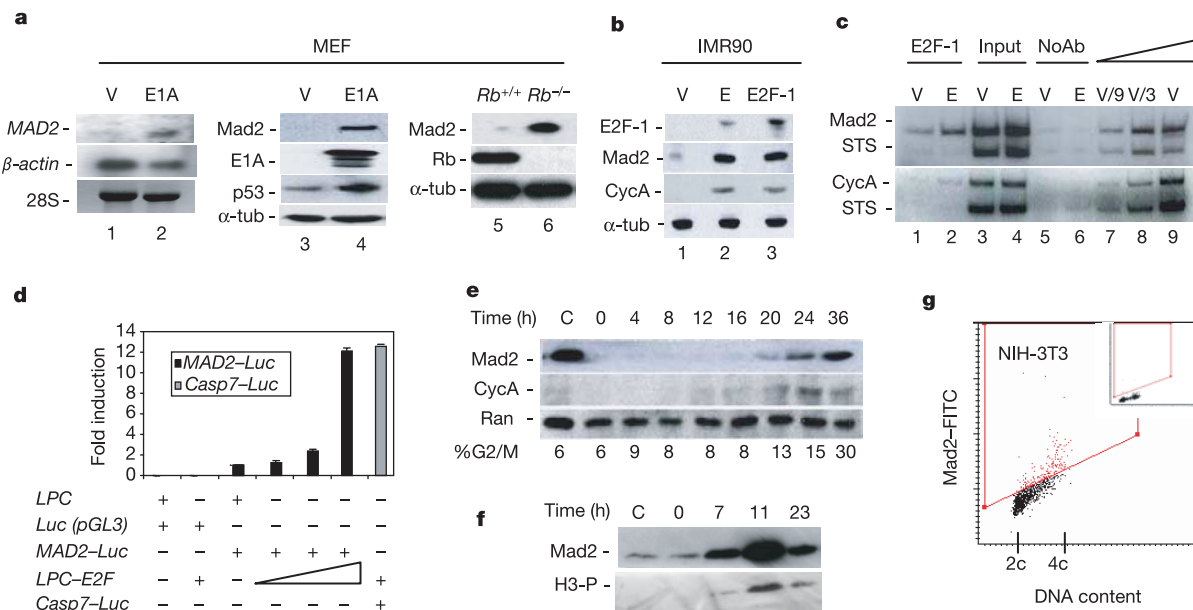


Figure 1 Mad2 is an E2F target. **a**, MAD2 mRNA (lane 1 and 2) and protein (lane 3 and 4) were assessed in vector (V) or E1A-expressing MEFs by northern blotting and immunoblotting, respectively. p53 induction indicates E1A activity and α-tubulin (α-tub), β-actin or 28S RNA serve as loading controls. Mad2 protein levels were also compared in wild-type (Rb^{+/+}) and Rb^{-/-} MEFs (lanes 5 and 6). **b**, Immunoblotting of the indicated proteins in IMR90 cells infected with vector- (V), E1A- (E) or E2F-1-expressing adenoviral constructs. Cyclin A (Cyc A) is an established E2F target. **c**, Chromatin immunoprecipitation¹⁵ using an E2F-1 specific antibody in vector- (V) and E1A (E)-expressing IMR90 cells. Duplex PCR amplifications were performed using primers surrounding the putative E2F sites in the Mad2 promoter or the known sites in the cyclin A promoter, along with primers capable of amplifying Cyclin A genomic sequence (STS) lacking E2F binding sites (see Supplementary Fig. S1a). Input corresponds to PCR

reactions containing 0.5% of total chromatin used in immunoprecipitation reactions. **d**, Relative luciferase signal in U-2OS cells co-transfected with empty vector (LPC) or E2F-1 (LPC-E2F) and a luciferase reporter with (MAD2-Luc) or without (Luc) Mad2 promoter sequences. E2F-1 plasmid concentration (triangle). The Casp7-luc reporter is the positive control. Values represent the mean ± s.e.m. ($n = 3$). **e**, Immunoblotting for Mad2 in control (C) or serum-starved IMR90 cells ($t = 0$), and at various times after serum addition. The percentages of cells in G2/M phase at the indicated time (shown below). **f**, Immunoblotting for Mad2 in IMR90 cells synchronized in S phase and released from a double-thymidine blockade. Histone H3-P is shown as a marker of mitosis. **g**, Mad2 expression by LSC in NIH-3T3 cells, where the gate represents the threshold considered positive for Mad2 based on the equivalent isotype control (inset). Cell populations of 2c and 4c DNA content are indicated.

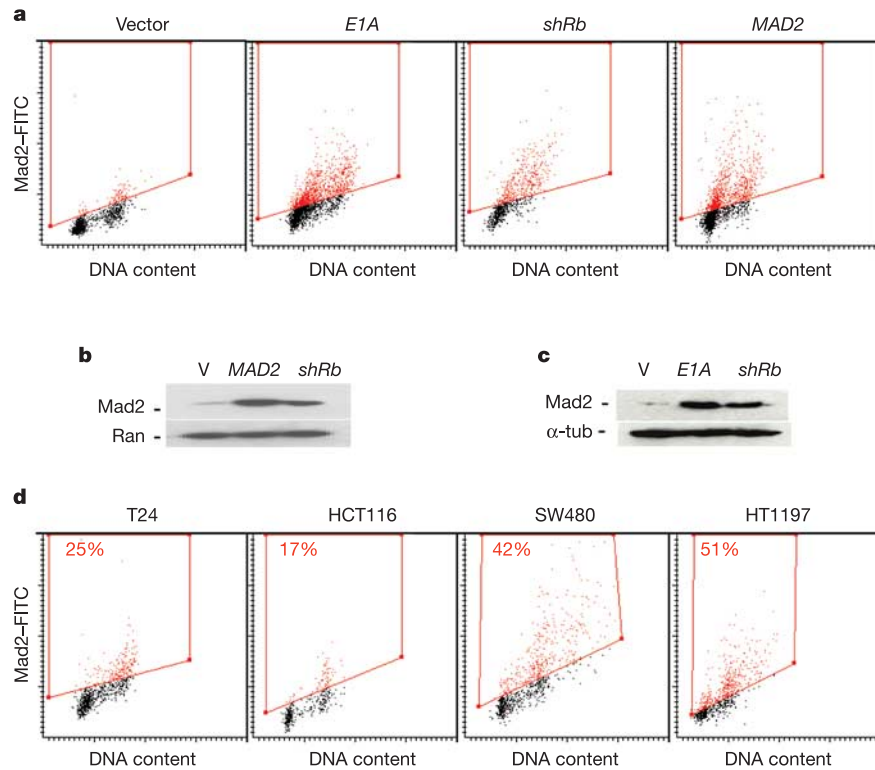


Figure 2 Deregulating the Rb/E2F pathway produces aberrant Mad2 expression. **a**, Biparametric analysis of Mad2 expression and DNA content by LSC in IMR90 fibroblasts expressing a vector control (Vector), *E1A*, *shRb* or *MAD2*. **b**, **c**, Western blot analysis of Mad2 in the cells described in part **a**. Ran and α -tubulin (α -tub) protein levels serve as

loading controls. **d**, LSC analysis of Mad2 expression and DNA content in T24, HCT116, HT1197 and SW480 tumour derived lines. Numbers indicate the total percentage of Mad2 positive cells.

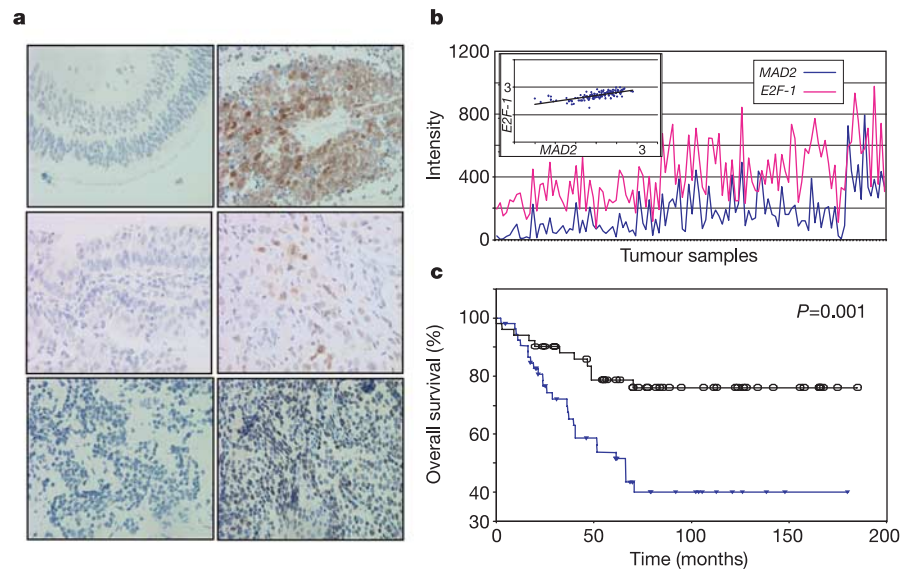


Figure 3 Abnormal Mad2 expression in human tumours. **a**, Immunohistochemistry for Mad2. Top: nearly undetectable Mad2 expression in normal retina (left) compared to high levels in retinoblastoma (right). Middle: an example of a Mad2-negative transitional cell carcinoma of the bladder (TCC, left) compared to a Mad2-positive invasive TCC (right). Bottom: a Mad2-negative (left) and positive (right) neuroblastoma. **b**, Comparison of *E2F-1* and *MAD2* transcript levels in a cohort of 106 neuroblastic tumours and 12

neuroblastoma cell lines²⁰. The inset represents the logarithm of *MAD2* mRNA expression (X axis) versus *E2F-1* (Y axis), indicating a near-linear correlation. **c**, Kaplan-Meier survival curves stratified by *MAD2* microarray-based mRNA levels, with prognostic significance by the log-rank test using median values as cut points. Neuroblastoma cases with low (black curve) and high (blue curve) *MAD2* mRNA levels are shown.

NIH-3T3 cell populations accumulated cells with greater than 4c DNA content during the course of cell division (Fig. 4a). These effects were exacerbated by treatment with nocodazole, which acutely engages the spindle checkpoint (data not shown). Analogous results were obtained using fluorescent *in situ* hybridization (FISH) to track specific chromosomes. The *E1A*-, *shRB*- and *MAD2*-expressing populations each accumulated cells with aberrant chromosome numbers (Fig. 4b, c). In contrast, IMR90 cells expressing a control vector retained normal chromosome content upon serial passaging (Fig. 4a–c, see Vector or V) and responded normally to nocodazole (data not shown). Chromosome counting of metaphase spreads confirmed that IMR90 and HCT116 cells with defective Rb or elevated Mad2 are frequently aneuploid (Fig. 4c). No centrosome amplification was detected in *E1A*- or *MAD2*-transduced fibroblasts by γ -tubulin (see Supplementary Fig. S3b) or β -tubulin immunofluorescence staining (spindle shape analysis; data not shown), suggesting that the observed mitotic defects are not due to the effects of cyclin E on centrosome duplication^{24,25}. However, as cyclin E is an E2F target, centrosome amplification may exacerbate the chromosomal instability associated with Rb loss. Nevertheless, misexpression of Mad2 is sufficient to produce this phenotype.

We next examined mitotic progression of individual cells in *E1A*-,

shRb-, *MAD2*- or vector-transduced populations co-expressing green fluorescent protein (GFP)-tagged histone (H2B–GFP), by phase contrast and fluorescence videomicroscopy. Whereas normal NIH-3T3 cells underwent an orderly and rapid mitosis, cells expressing *E1A* or *shRb* took an unusually long time to finish the process, displaying marked difficulties in completing cytokinesis (Fig. 5a–c; Supplementary Fig. S4a and S5, compare video A with C and D) and frequent defects in chromosome segregation (Fig. 5c; Supplementary Fig. S5, compare video B with G). Cells expressing Mad2 alone displayed similar defects (Fig. 5c; Supplementary Fig. S5, compare video A with E and F), although the chromosome segregation abnormalities were more pronounced, perhaps owing to the higher Mad2 levels that these populations expressed (Supplementary Fig. S5, compare video B with H). Similar phenotypes were also observed in *E1A*-, *shRb*- and *MAD2*-transduced HCT116 (Supplementary Fig. S5, videos I and J). In each cell population, the frequency of aberrant mitoses decayed after serial passaging, perhaps owing to lethality of cells with aberrant chromosome content and/or to cellular adaptation to elevated Mad2. Nevertheless, consistent with a prolonged mitosis, cells aberrantly expressing Mad2 displayed a delay in the degradation of securin and cyclin B (Supplementary Fig. S4b, c), whose sequential ubiquitination by the APC/cyclosome is inhibited by

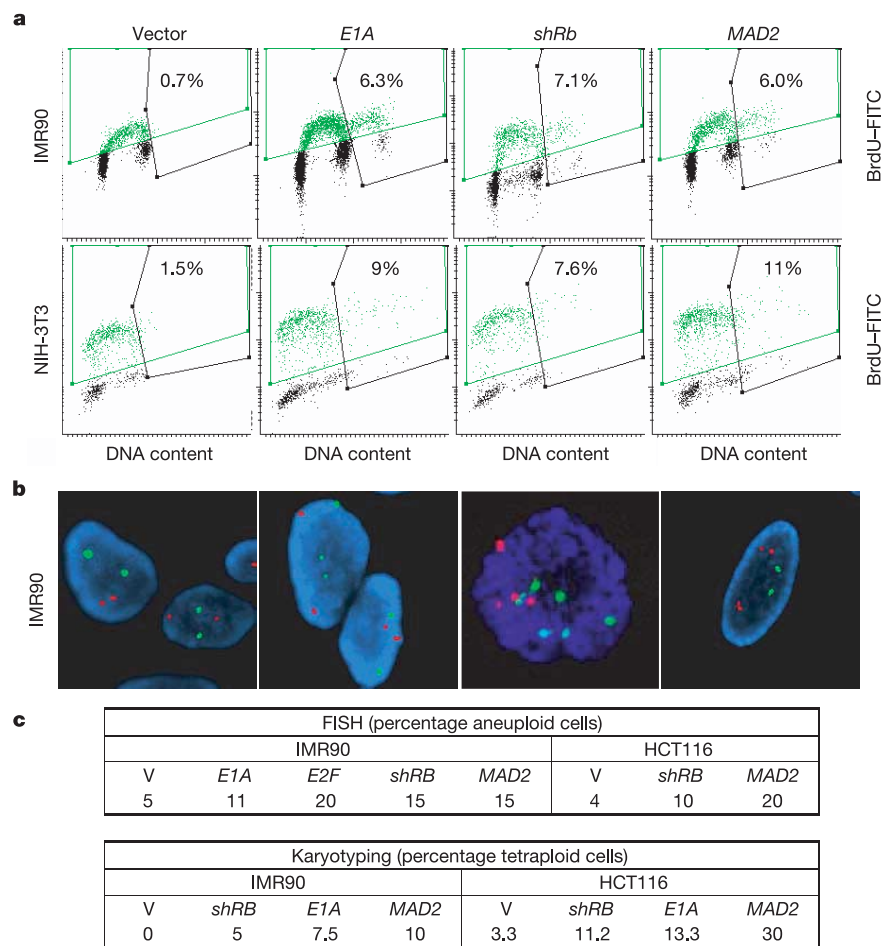


Figure 4 Cells with defective Rb or elevated Mad2 display chromosomal instability. **a**, Immunocytochemical detection of BrdU versus DNA content by LSC in asynchronous cultures of IMR90 (upper panels) or NIH-3T3 (lower panels) infected with *E1A*, *shRB* or *MAD2*. **b**, FISH analysis on IMR90 cells, with centromeric probes for chromosomes 1 and 17 and DAPI counterstaining. **c**, Quantification of ploidy alterations derived from *E1A*,

shRb or *MAD2* expression. FISH in IMR90 and HCT116 transduced cells (upper table). Shown are percentages of cells with abnormal number of positive signals for one or both probes. Percentage of infected cells exhibiting tetraploidy in metaphase spreads (lower table).

Mad2^{7,13} and required for chromosome segregation and cytokinesis, respectively^{6,7,26}.

Complete downregulation of Mad2 in cells with intact Rb produces aneuploidy after a premature anaphase²⁷. However, if Mad2 upregulation contributes to chromosome instability, we reasoned that a partial suppression of Mad2 in Rb defective cells would have the opposite effect—perhaps reducing the occurrence of aneuploid cells. Therefore, we used stable RNAi to partially suppress Mad2 in control, *E2F*- or *E1A*-transduced HCT116 cells. Consistent with a previous report²⁷, stable expression of two independent short-hairpin RNAs (*shMAD2*) in cells with low Mad2 levels further reduced Mad2 expression and increased the percentage of aneuploid cells (data not shown). In contrast, co-expression of *shMAD2* with *E2F* or *E1A* significantly reduced Mad2 to nearly normal levels (Fig. 5d, compare lanes 3 with 7 or 2 with 8) and reduced the fraction of aneuploid cells (Fig. 5e, f) without affecting *E1A* or *E2F*-1 levels (Supplementary Fig. S4d). Cells coexpressing *shMAD2* and either *E2F-1* or *E1A* were proliferating and showed similar S phase content compared to cells harbouring a control shRNA (data not shown). Therefore, although other Rb targets may contribute to

chromosome instability, aberrant Mad2 activity is important for this process.

We suggest that aberrant expression of Mad2 arising as a result of Rb pathway defects can produce a hyperactive spindle checkpoint, thereby altering the sequence of mitotic events and the accuracy of chromosome transmission. Persistent inactivation of the APC/cyclosome, mediated by Mad2 signalling, delays the degradation of securin and cyclin B. Moreover, overexpression of securin (Pds1, PTTG)—observed in many different tumour types—produces aneuploidy and displays transforming activity *in vivo* and *in vitro*²⁸. Interestingly, the behaviour of cells with aberrant Mad2 expression is reminiscent of cells treated for prolonged periods with nocodazole—following spindle checkpoint activation, these cells eventually adapt and enter the next cell cycle without a proper mitosis²⁹. Together, our results demonstrate that deregulated Mad2 expression contributes to mitotic alterations and chromosome instability in the context of an altered Rb/E2F pathway.

Our data illustrate how aneuploidy in human cancers can arise as a byproduct of the Rb pathway defects that accompany tumorigenesis. During normal cell proliferation, the cell cycle is hardwired to

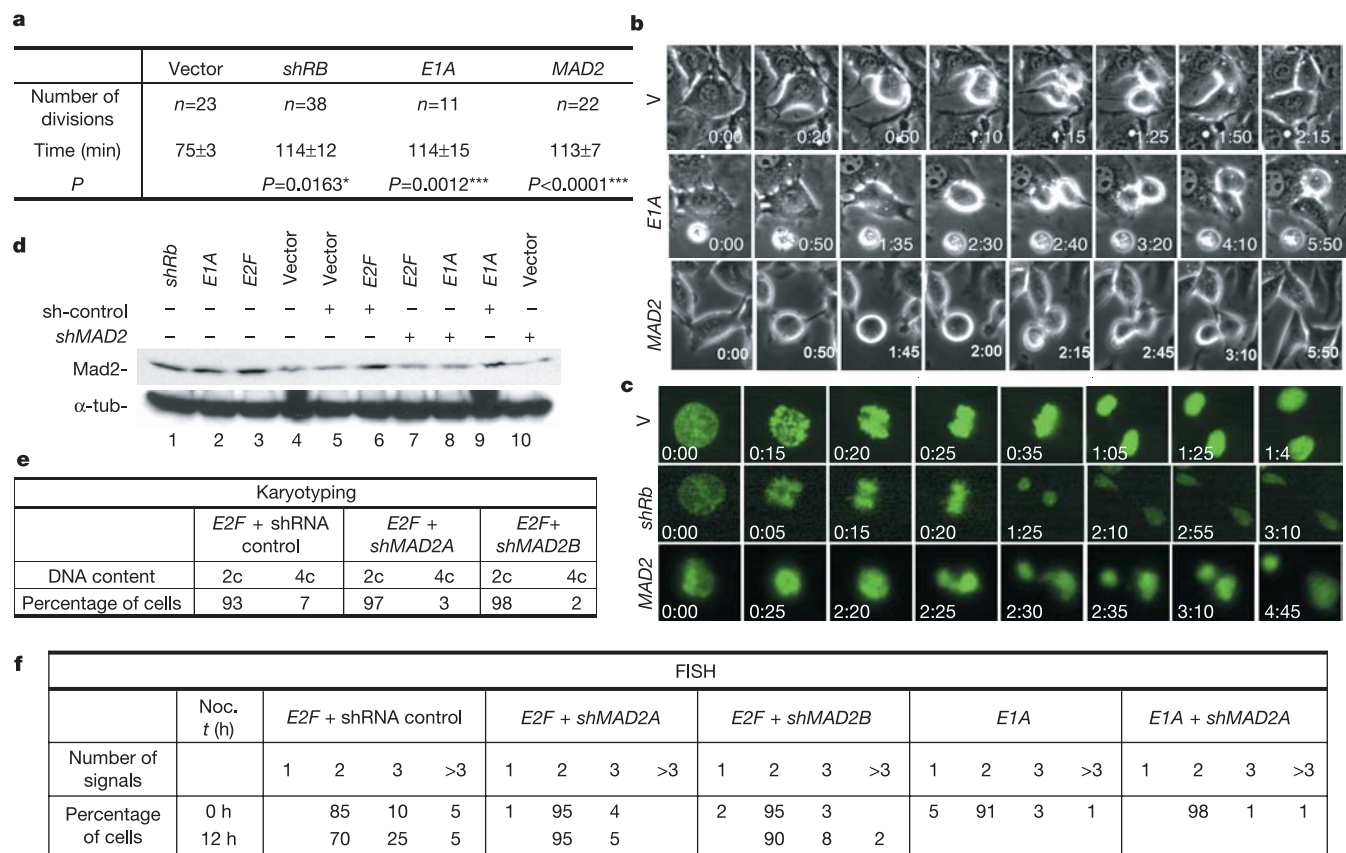


Figure 5 Aberrant Mad2 expression contributes to the mitotic defects associated with Rb loss. **a**, Average time needed to complete mitosis in NIH-3T3 cells expressing control vector, *shRb*, *E1A* or *MAD2* assessed from prometaphase (disruption of the nuclear envelope and chromatin condensation) to completed cytokinesis and reattachment of daughter cells. *n*, number of cells analysed. *P* values determined by Student's *t*-test. **b**, Time course of a normal mitotic division occurring in vector-transduced NIH-3T3 cells (v). Illustrative examples of prolonged mitoses occurring in cells expressing *E1A*, *shRb* or *MAD2* (*E1A*- and *MAD2*-transduced NIH-3T3 cells shown). **c**, A fluorescence videomicroscopy series of H2B-GFP transfected cells showing normal chromosome segregation (vector-transduced NIH-3T3, v) and representative examples of the slow and abnormal chromosomal excision patterns observed in cells expressing *shRb*, *E1A*

or *MAD2* (*shRb* and *MAD2* shown). **d**, Immunoblotting for Mad2 in HCT116 cells expressing *E2F*-, *E1A*- or Vector together with either a *MAD2* hairpin (*shMAD2*) or a hairpin control (*sh-control*) as indicated. **e**, **f**, Cytogenetic analyses by karyotyping (**e**) or FISH (**f**) of HCT116 cells co-expressing *E2F* or *E1A* with a *MAD2* shRNA (two sequence-independent hairpins, *shMAD2A* and *shMAD2B*) or a shRNA control. ≥200 interphase nuclei were analysed by FISH. Values were compared for statistical significance using the exact test of Fisher: *P* = 0.015 and *P* < 0.0001, comparing *E2F*-transduced cells coinfecting with *shMAD2A* or shRNA control, at 0 and 12 h nocodazole (Noc.); *P* = 0.006 and *P* = 0.0007, for *shMAD2B* versus shRNA control, at 0 and 12 h respectively.

the spindle checkpoint through E2F, presumably to ensure that Mad2 expression is tightly controlled. As a consequence, mutations that disrupt Rb function produce deregulated E2F activity leading to inappropriate proliferation, but also aberrant Mad2 expression and, eventually, aneuploidy. Hence, our results provide one explanation for why loss-of-function mutations in spindle checkpoint genes are uncommon in human cancers^{11,12}. Our model also implies that aneuploidy is not necessarily a selected trait of cancer cells but can be an inadvertent consequence of defects that uncouple mitotic control from normal cell cycle progression. Nevertheless, this byproduct of Rb loss might subsequently increase the mutation rate, thereby fuelling further tumour evolution. □

Methods

Cells, gene transfer and drug treatment

Human tumour cell lines (T24, HT1197, HCT116, SW480 and U-2OS), normal human fibroblasts (IMR90) and murine immortal fibroblasts (NIH-3T3) were obtained from the ATCC (American type culture collection, Rockville). Primary MEFs were isolated and cultured as described¹⁵. Cells were infected with high-titre recombinant retroviruses expressing *E1A* (LPC-12S), *E2F-1* (LPC-E2F1), *shRb* (MSCV-shRB)¹⁶, *shMAD2* (MSCV-shMAD2A or B)²⁷, *shmurMAD2* (shRNA control, MSCV-shmurMAD2)²⁷ or *MAD2* cDNA (Babe-MAD2) and selected with puromycin (2 µg ml⁻¹)¹⁵. Adenoviruses *E1A*, *E2F-1* and *GFP* control were used at a multiplicity of infection of 100 PFU cell⁻¹ as described¹⁵. Nocodazole (Sigma) was used at 200 nM and 400 nM for normal fibroblast or tumour cell lines, respectively.

Protein and mRNA expression

Protein expression was assessed by immunoblotting using 30 µg of total cell lysate¹⁵. Blots were probed with antibodies directed against hsmad2 (clone 48, BD Biosciences), E1A (M58, BD Biosciences), p53 (CM5, Novocastra), E2F-1 (KH95, Santa Cruz), Cyclin A (BF683, Santa Cruz), Ran (sc-1156, Santa Cruz), Pds1 (ab-1, DCS-280, Neomarkers), cyclin B1 (05-373, Upstate), phospho-histone3 (06-570, Upstate), APC2 (ab-1, Neomarkers) or α -tubulin (B-5-1-2, Sigma). Anti-mouse, anti-rabbit (Amersham) or anti-goat horseradish peroxidase (HRP) (Sigma) were used as secondary antibodies and proteins were visualized using an ECL detection system (Amersham). mRNA levels were assessed by northern blotting using 10 µg of total RNA¹⁵.

LSC and FACS

Cells grown as monolayers on microscope-chamber slides were washed, fixed and incubated with the respective antibodies³⁰: hsmad2 (BD Biosciences); α -tubulin FITC conjugated (clone DM 1A) and γ -tubulin (Sigma); and anti-Rb (Pharmingen, BD Laboratories). Mouse IgG served as an isotype negative control. Cells were washed and incubated with FITC-conjugated rabbit anti-mouse F(ab')₂ or anti-rabbit (DAKO), then washed and incubated with 5 µg ml⁻¹ of propidium iodide (PI; Molecular Probes) and 0.1% RNase A (Sigma) in PBS before measurement by LSC (CompuCyte) as indicated³⁰. A minimum of 3,000 cells were analysed per slide as previously described³⁰.

Cell cycle analysis after low-serum synchronization of IMR90 cells was performed as previously described¹⁵. Mitotic index was quantified by measuring MPM2 expression (anti-MPM2, Upstate Biotechnology) versus DNA content (PI) by FACS²⁷. For BrdU incorporation assays, slide-grown cells were incubated in 10 µM BrdU (Sigma) for 1 h 30 min (human cell lines) or 3 h (normal fibroblasts) immediately before collection. Cells were fixed with 80% ethanol overnight and incubated with FITC-conjugated anti-BrdU antibody after denaturing the DNA, following the manufacturer's instructions (Pharmingen, BD Laboratories). BrdU incorporation was measured by LSC, with the equivalent FITC-conjugated isotypic control as a baseline.

Chromosome analyses

Bicolour FISH was performed using probes for chromosomes 1 and 17 centromere-specific alphoid region (Vysis)²⁰. The number of hybridization signals for each probe was assessed in a minimum of 200 interphase nuclei with strong and well-delineated contours. Chromosome analysis of cells in metaphase was performed as described²⁷.

Immunohistochemistry

Sections were deparaffinized, treated with 1% H₂O₂, immersed in boiling 10 mM citrate buffer for 15 min and incubated in 10% normal horse serum for 30 min at room temperature. Anti-hsmad2 (Transduction Laboratories, BD Laboratories; dilution 1:100) and anti-human Ki-67 (MIB-1, DAKO; dilution 1:1000) antibodies were used. Samples were then incubated with biotinylated anti-mouse IgG at 1:500 dilution (Vector Laboratories) followed by avidin-biotin peroxidase complexes (1:25; Vector Laboratories) for 30 min. Diaminobenzidine was used as the chromogen, and hematoxylin as the nuclear counterstain.

Human tumour samples and gene expression analysis

Twelve cases of retinoblastoma and 95 bladder tumours on a tissue array were analysed by immunohistochemistry. In addition, 12 neuroblastoma cell lines and 106 neuroblastic tumours were studied²⁰. Total RNA was extracted and cDNA was labelled and hybridized to Affymetrix human U95 oligonucleotide arrays as described²⁰. Comparisons of *E2F-1*, *MAD2* and *MYCN* expression levels were performed by Kendall's *tau* correlation test.

Kaplan–Meier survival curves were generated by using the SPSS program (SPSS). *E2F-1*, *MYCN* and *MAD2* expression were tested for prognostic significance ($P < 0.05$) by the log-rank test using median values as cut points. Protein expression was validated on a tissue microarray containing 50 neuroblastoma cases.

Received 15 April; accepted 6 July 2004; doi:10.1038/nature02820.

- Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
- Albertson, D. G., Collins, C., McCormick, F. & Gray, J. W. Chromosome aberrations in solid tumors. *Nature Genet.* **34**, 369–376 (2003).
- Dyson, N. The regulation of E2F by pRB-family proteins. *Genes Dev.* **12**, 2245–2262 (1998).
- Jallepalli, P. V. & Lengauer, C. Chromosome segregation and cancer: cutting through the mystery. *Nature Rev. Cancer* **1**, 109–117 (2001).
- Cahill, D. P. *et al.* Mutations of mitotic checkpoint genes in human cancers. *Nature* **392**, 300–303 (1998).
- Cleveland, D. W., Mao, Y. & Sullivan, K. F. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell* **112**, 407–421 (2003).
- Yu, H. Regulation of APC-Cdc20 by the spindle checkpoint. *Curr. Opin. Cell Biol.* **14**, 706–714 (2002).
- Kalitsis, P., Earle, E., Fowler, K. J. & Choo, K. H. *Bub3* gene disruption in mice reveals essential mitotic spindle checkpoint function during early embryogenesis. *Genes Dev.* **14**, 2277–2282 (2000).
- Dobles, M., Liberal, V., Scott, M. L., Benezra, R. & Sorger, P. K. Chromosome missegregation and apoptosis in mice lacking the mitotic checkpoint protein Mad2. *Cell* **101**, 635–645 (2000).
- Michel, L. S. *et al.* MAD2 haplo-insufficiency causes premature anaphase and chromosome instability in mammalian cells. *Nature* **409**, 355–359 (2001).
- Cahill, D. P. *et al.* Characterization of MAD2B and other mitotic spindle checkpoint genes. *Genomics* **58**, 181–187 (1999).
- Hernando, E. *et al.* Molecular analyses of the mitotic checkpoint components *hMAD2*, *hBUB1* and *hBUB3* in human cancer. *Int. J. Cancer* **95**, 223–227 (2001).
- Wassmann, K. & Benezra, R. Mad2 transiently associates with an APC/p55Cdc complex during mitosis. *Proc. Natl Acad. Sci. USA* **95**, 11193–11198 (1998).
- Ren, B. *et al.* E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev.* **16**, 245–256 (2002).
- Nahlé, Z. *et al.* Direct coupling of the cell cycle and cell death machinery by E2F. *Nature Cell Biol.* **4**, 859–864 (2002).
- Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703–716 (2003).
- Sanchez-Carbayo, M. *et al.* Molecular profiling of bladder cancer using cDNA microarrays: defining histogenesis and biological phenotypes. *Cancer Res.* **62**, 6973–6980 (2002).
- Arber, N. *et al.* Antisense to cyclin D1 inhibits the growth and tumorigenicity of human colon cancer cells. *Cancer Res.* **57**, 1569–1574 (1997).
- Dalbagni, G., Presti, J., Reuter, V., Fair, W. R. & Cordon-Cardo, C. Genetic alterations in bladder cancer. *Lancet* **324**, 469–471 (1993).
- Alaminos, M. *et al.* Genome-wide analysis of gene expression associated with *MYCN* in human neuroblastoma. *Cancer Res.* **63**, 4538–4546 (2003).
- Schuldiner, O. & Benvenisty, N. A DNA microarray screen for genes involved in c-MYC and N-MYC oncogenesis in human tumors. *Oncogene* **20**, 4984–4994 (2001).
- Van't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
- Lentini, L., Pipitone, L. & Di Leonardo, D. Functional inactivation of pRB results in aneuploid mammalian cells after release from a mitotic block. *Neoplasia* **4**, 380–387 (2002).
- Spruck, C. H., Won, K. A. & Reed, S. I. Deregulated cyclin E induces chromosome instability. *Nature* **401**, 297–300 (1999).
- Okuda, M. *et al.* Nucleophosmin/B23 is a target of CDK2/cyclin E in centrosome duplication. *Cell* **103**, 127–140 (2000).
- Stemmann, O., Zou, H., Gerber, S. A., Gygi, S. P. & Kirschner, M. W. Dual inhibition of sister chromatid separation at metaphase. *Cell* **107**, 715–726 (2001).
- Michel, L. *et al.* Complete loss of the tumor suppressor MAD2 causes premature cyclin B degradation and mitotic failure in human somatic cells. *Proc. Natl Acad. Sci. USA* **101**, 4459–4464 (2004).
- Zou, H., McGarry, T. J., Bernal, T. & Kirschner, M. W. Identification of a vertebrate sister-chromatid separation inhibitor involved in transformation and tumorigenesis. *Science* **285**, 418–422 (1999).
- Lanni, J. S. & Jacks, T. Characterization of the p53-dependent postmitotic checkpoint following spindle disruption. *Mol. Cell. Biol.* **18**, 1055–1064 (1998).
- Darzynkiewicz, Z., Gong, J., Juan, G., Ardel, B. & Traganos, F. Cytometry of cyclin proteins. *Cytometry* **25**, 1–13 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Narita for providing the shRb construct, M. Lu for the bladder tissue microarray, S. Menendez for technical assistance and A. Kel for the use of the SiteScan program. We thank the Molecular Cytology and the Flow Cytometry Core Laboratories at Memorial Sloan-Kettering and P. McCloskey from CSHL flow cytometry facility for technical assistance. Also, we thank S. Gangadharan, R. Dickens, S. Gonzalez, N. Abumrad, B. Stillman and P. P. Pandolfi for reading the manuscript. We also thank all the members of the Lowe and Cordon-Cardo laboratories for discussions. This work was supported by program project grants from the NCI (S.W.L. and C.C.C.), Clinical Investigator Research Development Awards (L.M. and R.B.), DOD-Breast Cancer Research Program fellowship award (Z.N.), a Fundacion Caja Madrid-CNIO-MSKCC fellowship (E.H. and M.A.), a Fellowship from the Spanish Ministry of Education, Culture and Sports (E.D.R.), a US NCI postdoctoral training grant (M.T.H.) and gifts from the Laurie Strauss Leukemia Foundation and the Herbert J. Siegel philanthropic fund.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.W.L. (lowe@cshl.edu).

Structure of the molybdopterin-bound Cnx1G domain links molybdenum and copper metabolism

Jochen Kuper¹*, Angel Llamas¹, Hans-Jürgen Hecht², Ralf R. Mendel¹ & Günter Schwarz¹

¹Department of Plant Biology, Technical University, Spielmannstrasse 7, D-38106 Braunschweig, Germany

²German Research Center for Biotechnology, Mascheroder Weg 1, D-38124 Braunschweig, Germany

* Present address: EMBL Hamburg outstation, DESY, D-22603 Hamburg, Germany

The molybdenum cofactor is part of the active site of all molybdenum-dependent enzymes¹, except nitrogenase. The molybdenum cofactor consists of molybdopterin, a phosphorylated pyranopterin², with an ene-dithiolate coordinating molybdenum. The same pyranopterin-based cofactor is involved in metal coordination of the homologous tungsten-containing enzymes found in archaea³. The molybdenum cofactor is synthesized by a highly conserved biosynthetic pathway⁴. In plants, the multidomain protein Cnx1 catalyses the insertion of molybdenum into molybdopterin. The Cnx1 G domain (Cnx1G), whose crystal structure has been determined in its apo form, binds molybdopterin with high affinity and participates in the catalysis of molybdenum insertion. Here we present two high-resolution crystal structures of Cnx1G in complex with molybdopterin and with adenylated molybdopterin (molybdopterin-AMP), a mechanistically important intermediate. Molybdopterin-AMP is the reaction product of Cnx1G and is subsequently processed in a magnesium-dependent reaction by the amino-terminal E domain of Cnx1 to yield active molybdenum cofactor. The unexpected identification of copper bound to the molybdopterin dithiolate sulphurs in both structures, coupled with the observed copper inhibition of Cnx1G activity, provides a molecular link between molybdenum and copper metabolism.

As a catalytic centre, the molybdenum cofactor (Moco) has important roles in the global carbon, sulphur and nitrogen cycles⁵. Molybdenum enzymes are important for diverse metabolic processes, such as sulphur detoxification and purine catabolism in mammals⁶, and nitrogen assimilation and phytohormone synthesis in plants⁷. Deficiency of Moco in humans results in the pleiotropic loss of all molybdenum enzyme activities⁸, resulting in neurological abnormalities⁹ and early childhood death. In Moco biosynthesis⁴, two intermediates, the sulphur-free precursor Z and molybdopterin (MPT), are synthesized before a final step in which a molybdenum atom is attached to one (pro- and eukaryotes) or two (prokaryotes) MPT ene-dithiolates. The active site of Cnx1G, which catalyses one step in the plant molybdenum insertion reaction (step 3), has been mapped to a large surface depression with one side essential for MPT binding and the other essential for catalysis¹⁰. The other domain, Cnx1E, is thought to participate in molybdenum activation and insertion^{11,12}.

To examine the active site of Cnx1G, we have determined the crystal structure of wild-type Cnx1G in complex with MPT (Cnx1G-MPT; Supplementary Table 1). Cnx1G-MPT formed *in vivo* was purified from *Escherichia coli* and crystallized anaerobically. The 2F_o-F_c density in the 1.6 Å structure of Cnx1G-MPT clearly indicated bound MPT in the 'left' side of the surface depression (Fig. 1a, e). All MPT atoms were defined by electron density (Fig. 1c). The absence of a bond between C1' and the dithiolate sulphur is probably due to radiation damage, because the C1'-S distance (1.79 Å) is within the limits of a Moco C1'-S bond¹³. The strong difference density between both dithiolate sulphurs is

discussed below. Two residues, Thr 542 and Ser 573, have been shown to be important for MPT binding¹⁰. These residues are involved in the hydrogen-bonding network to MPT in the Cnx1G-MPT structure (Fig. 1e and Supplementary Fig. 1).

The second structure, at 1.45 Å resolution, of the fully active Ser583Ala variant of Cnx1G (hereafter referred to as Ser583Ala) shows the same complex as Cnx1G-MPT (Fig. 1b, e). However, additional density in the 'right' depression, which was previously proposed to be a catalytic site essential for molybdenum insertion¹⁰, is an adenosine bound by means of a pyrophosphate bond to MPT (Fig. 1b, d). Only the ribose atoms C4' and C5' and the C5' hydroxyl group show weak or no electron density. The unexpected observation of adenylated MPT (MPT-AMP) has mechanistic implications for the molybdenum insertion reaction. All residues identified in previous studies^{10,14} that are important for the catalytic function of Cnx1G (Ser 476, Asp 486, Asp 515 and Asp 548) are in close proximity to the adenosine or the pyrophosphate bond (Fig. 1e), and the electrostatic surface potential closely matches the charge of the MPT-AMP ligand (Fig. 1f). Although mature prokaryotic Moco may carry nucleotides, the adenylation of MPT has not been reported so far. AMP activation seems to be a common feature in Moco biosynthesis, because the sulphur-transfer reaction during MPT synthesis is also dependent on a ubiquitin-like adenylation reaction¹⁵.

The structural reason for the observation of MPT-AMP in the Ser583Ala structure (Supplementary Fig. 2) rests on the hydrophobic serine to alanine substitution, which causes, as compared

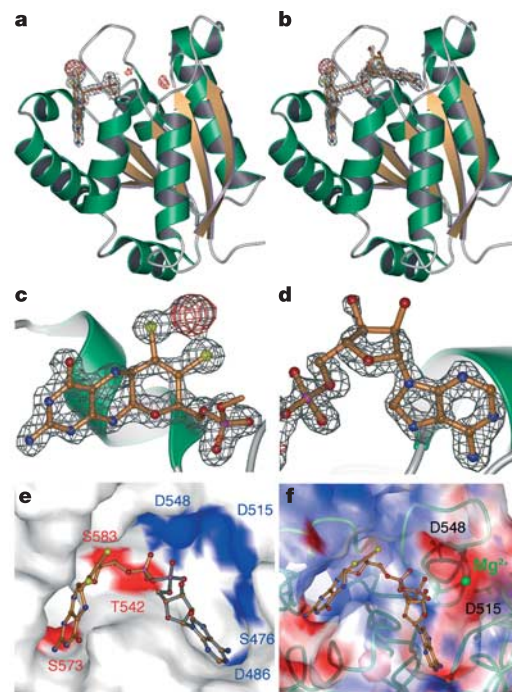


Figure 1 Structure of the Cnx1G-MPT and Ser583Ala-MPT-AMP complexes.

a, b, Ribbon representation of Cnx1G (**a**) and Ser583Ala (**b**). α -Helices and β -strands are shown in green and orange, respectively. Bound MPT and MPT-AMP are shown in ball-and-stick notation and are covered by the 2F_o-F_c electron density contoured at 1.0 σ . The difference density (F_o-F_c) is shown in red (3.5 σ). **c, d**, Close-up views of the bound MPT (**c**) and AMP (**d**) regions in the Ser583Ala structure. **e, f**, Molecular surface of Ser583Ala with bound MPT-AMP showing, respectively, residues important for MPT binding (red) and catalysis^{12,14} (blue), and the electrostatic surface potential with ± 11 kT as borders for electropositive (blue) and negative (red) regions calculated with SYBIL and GRASP³⁰. The catalytically important residues Asp 515 and Asp 548 are highlighted. The position of Mg²⁺ found in the homologous MoeA²² domain is superimposed.

with Cnx1G–MPT, a 0.4 Å shift of the pterin backbone and surrounding residues (601–607). MPT is surface exposed in both crystal structures, in contrast to the deeply buried cofactor in molybdenum- or tungsten-containing enzymes^{2,13,16,17} (Supplementary Fig. 3). The pyranopterin nature of Moco, identified in crystal structures of these enzymes, is now shown for the biosynthetic intermediate MPT, which is consistent with the spectroscopic characterization of precursor Z¹⁸, another intermediate in Moco biosynthesis. We therefore conclude that the ring-opened dihydro form of MPT in bacterial nitrate reductase occurs after cofactor insertion¹⁹.

We investigated whether MPT–AMP is the substrate or the reaction product of Cnx1G. The MPT content of Cnx1G and Ser583Ala was similar to that of Asp515His¹⁴, a variant with wild-type-like MPT binding but complete loss of Moco-synthetic activity (Fig. 2a). The amount of bound MPT–AMP was higher in Ser583Ala than in Cnx1G, and there was no detectable MPT–AMP in Asp515His, clearly showing that Cnx1G functions in the adenylation of MPT (Fig. 2a). Synthesis of Moco was assayed by the reconstitution of apo-nitrate reductase in desalted *Neurospora crassa nit-1* protein extracts^{14,20} (Fig. 2b). With excess molybdate, all three proteins (Cnx1G, Ser583Ala and Asp515His) formed active Moco. An unknown low molecular mass compound¹⁴ needed for Cnx1G-dependent Moco synthesis in the absence of external molybdate was efficiently replaced by 100 μM MgCl₂.

In Ser583Ala, the Mg²⁺-dependent activity was increased, owing to the larger amount of bound MPT–AMP, whereas Asp515His showed no activity because it cannot synthesize MPT–AMP. Because no additional molybdate was needed for this Mg²⁺-dependent activity, it seems that trace amounts of protein-bound molybdate present in completely desalted *nit-1* extract are sufficient for Moco synthesis. As a loss of molybdenum insertion is observed on mutation of Cnx1G or Cnx1E¹¹, we investigated Cnx1E in the Cnx1G-dependent conversion of MPT–AMP to Moco (Fig. 2c). Cnx1E increased Moco synthesis from Cnx1G and Ser583Ala, but only in the presence of MgCl₂. The fact that Cnx1E is not exclusively needed in the *nit-1* assay for Moco synthesis can be explained by the presence of a Cnx1 homologue in the *nit-1* extract. We therefore conclude that Cnx1E cleaves MPT–AMP, the product of Cnx1G, and transfers molybdenum to the MPT dithiolate to yield Moco.

Synthesis and cleavage of the MPT adenylate by Cnx1G and Cnx1E, respectively, agree well with the structural similarity observed between the corresponding domains of their bacterial

homologues MogA and MoeA²¹. In MoeA, a bound Mg²⁺ ion coordinated by six water molecules has been identified²². When superimposing the MoeA domain III, which comprises most of the active site, onto the similarly folded Ser583Ala–MPT–AMP structure (Supplementary Fig. 4), the MoeA-bound Mg²⁺ ion ends up above the pyrophosphate of MPT–AMP between Asp 515 and Asp 548 (Fig. 1f). These residues, which are also conserved in MoeA (Asp 228 and Asp 258) and are directly involved in coordinating the Mg²⁺-water cluster, are catalytically important¹⁴ and might therefore participate in the adenylation of MPT by coordinating the cation.

Both structures (Fig. 1a–c) show strong difference density between both MPT dithiolate sulphurs. The presence of molybdenum was ruled out by comparing the anomalous scattering

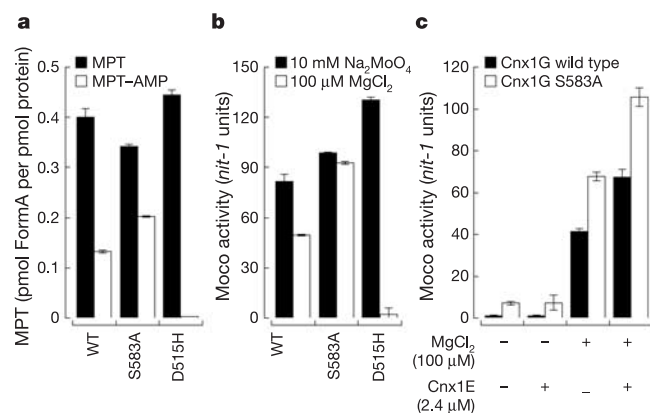


Figure 2 Mg²⁺- and Cnx1E-dependent Moco synthesis by Cnx1G–MPT complexes. **a**, Content of MPT and MPT–AMP in 200 pmol of Cnx1G, Ser583Ala and Asp515His. **b**, **c**, *In vitro* synthesis of Moco using the MPT or MPT–AMP complexes of Cnx1G, Ser583Ala and Asp515His in the presence of molybdate and MgCl₂ (**b**, **c**) and Cnx1E (**c**).

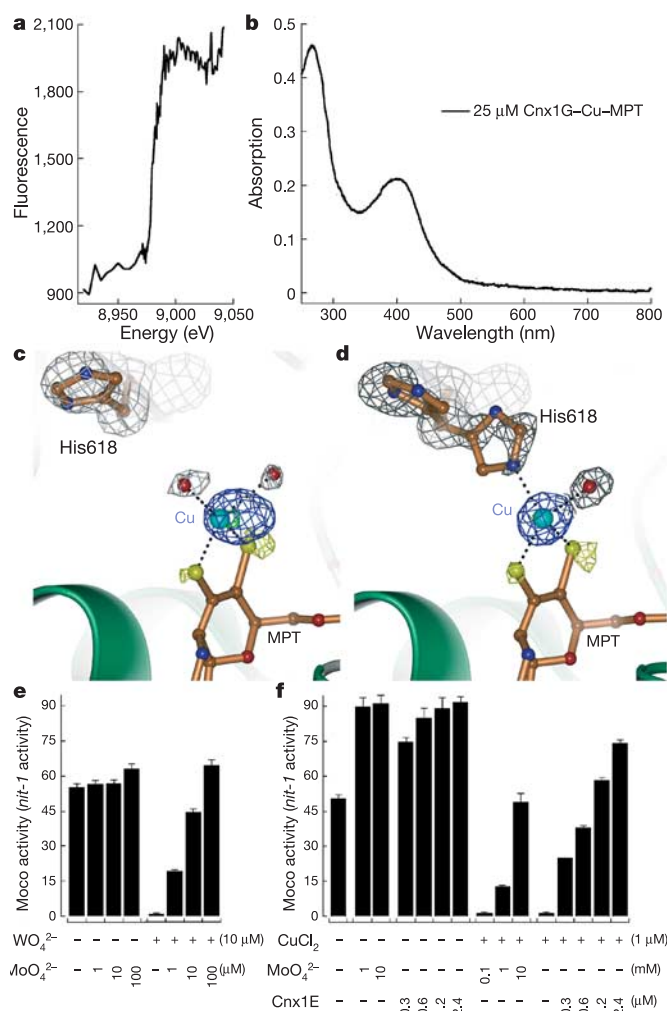


Figure 3 Cu–MPT dithiolate complex. **a**, Fluorescence scan of the Cnx1G crystal at the copper absorption edge. **b**, Ultraviolet–visible spectrum of the Cnx1G–MPT complex at 25 μM. **c**, Anomalous Fourier maps (coloured) of Cnx1G surrounding the copper site and 2F_o–F_c maps (grey) for non-dithiolate ligands. Broken lines indicate the bonds between copper and coordinating atoms. Anomalous Fourier maps were calculated from diffraction data of Cnx1G recorded before (1.377 Å, blue) and after (1.385 Å, green) the copper edge (4.5σ). **d**, Anomalous Fourier maps (depicted as in **c**) of the Ser583Ala variant derived from diffraction data recorded at 1.05 Å (blue, 5.0σ) and 1.75 Å (yellow, 3.0σ). **c**, **d**, A third anomalous Fourier map obtained from the 1.75 Å data is shown in yellow (3.0σ) and highlights the sulphur atoms. **e**, **f**, *In vitro* synthesis of Moco using the Cnx1G–MPT complex in the presence of 100 μM MgCl₂ at different concentrations of molybdate or Cnx1E with or without 10 μM tungstate (**e**) and 1 μM CuCl₂ (**f**).

recorded at 1.05 Å and 1.75 Å wavelengths (Supplementary Table 1); however, the anomalous signals of the MPT dithiolate sulphurs were assigned (Fig. 3c, d). The copper absorption edge was identified (Fig. 3a), and two additional data sets were collected (Supplementary Table 1) before (1.377 Å) and after (1.385 Å) the edge. A strong anomalous copper signal contoured at 4.5σ can be seen between the dithiolate at 1.377 Å, and this signal is almost absent at the same contour level at 1.385 Å (Fig. 3c). Although only 1.05 Å and 1.75 Å data were collected for Ser583Ala, the reduced signal at 1.75 Å shows that copper is bound in both structures (Fig. 3d) with similar bond distances of 2.26–2.28 Å to the MPT dithiolate (Supplementary Fig. 1). In comparison, all Mo–S distances are 2.41 Å in sulphite oxidase¹³.

Either two water molecules (Cnx1G) or one water molecule and His 618 (Ser583Ala) form two additional copper ligands (Fig. 3c, d). The conformational change of His 618 is due to the above-mentioned shift of the pterin backbone in Ser583Ala, which reduces the distance between copper and His 618, resulting in a tetrahedral coordination. The Cnx1G–Cu–MPT complex has a strong absorption at 400 nm (Fig. 3b). Molybdenum enzymes containing only Moco as a cofactor, such as plant sulphite oxidase¹⁷, show absorption maxima at 350 and 480 nm, owing to the ene-dithiolate-to-molybdenum and the cysteine-to-molybdenum charge transfer bonds, respectively¹⁷.

Moco synthesis with Cnx1G–Cu–MPT showed a loss of activity in the presence of 10 μ M tungstate, which is known to be a strong competitor. This inhibition was efficiently suppressed by equimolar amounts of molybdate (Fig. 3e), showing the preferred and catalysed incorporation of molybdenum into Cu–MPT. A similar inhibition of Moco synthesis was found with 1 μ M CuCl₂ (Fig. 3f). In contrast to tungstate, copper inhibition was rescued only by 10 mM molybdate, which is considered to cause non-catalysed molybdenum ligation, as also observed for inactive Cnx1G variants (for example, Asp515His)¹⁴. Copper inhibition of Moco synthesis can be explained either by increased stability of the Cu–MPT complex or by inhibition of the Mg²⁺-dependent molybdenum insertion reaction. The latter is supported by the suppression of copper inhibition with equimolar amounts of Cnx1E (Fig. 3f) and is in line with the known copper inhibition of pyrophosphatases²³. Our finding implies that Moco deficiency might occur when cellular

copper concentrations are increased, as seen in individuals affected with Wilson's disease²⁴, where copper accumulates in liver and brain, resulting in severe damage to both organs. The purification of an *in-vivo*-formed Cu–MPT complex on expression in molybdate-rich medium suggests, however, that copper protects the reactive dithiolate before the insertion of molybdenum.

In summary, we have shown that the final step of Moco biosynthesis is dependent on or associated with at least two unexpected processes: the adenylation of MPT, and copper–dithiolate complex formation (Fig. 4). Cnx1G synthesizes MPT–AMP, which is required for the subsequent Mg²⁺-dependent molybdenum-insertion reaction catalysed by Cnx1E. The observed inhibition of Moco synthesis by copper suggests a link to copper transport disorders, where some of the symptoms might be attributed to Moco deficiency. However, this copper–molybdenum antagonism is opposite to the previously known molybdenum–copper antagonism, where increased molybdate causes copper deficiency owing to the formation of copper-chelating thiomolybdates²⁵. □

Methods

Expression and purification of recombinant proteins

Mutagenesis of pQE60cnx1g628 (ref. 13) and expression were done as described¹⁴ with the following changes. To boost MPT synthesis to obtain Cnx1G highly saturated with MPT, we expressed RK5206/pQE60cnx1g628 for 30 h at 22 °C in LB-medium containing 100 μ M sodium molybdate and 15 μ M isopropyl β -thiogalactoside. Cnx1G–MPT complex was first purified by using nickel–nitrilotriacetic acid superflow matrix (Qiagen; 1 ml per litre of culture) as described¹⁴, but under anaerobic conditions. Further separation of MPT-bound from apo-Cnx1G variants has been achieved by cation exchange chromatography on an HR5/5 MonoS column (Amersham) equilibrated in 50 mM acetate buffer (pH 5.0) plus 50 mM NaCl. MPT-bound Cnx1G eluted at about 300 mM NaCl. The MPT-containing protein fractions were pooled, concentrated to 20 mg ml^{−1} and stored in aliquots in liquid nitrogen.

Crystallization and structure determination

Cnx1G and Ser583Ala were crystallized at 22 °C by vapour diffusion under anaerobic conditions using sitting-drop plates with 2–4 μ l of protein solution and 2 μ l of reservoir solution (2.8–3.6 M sodium formate). Crystals appeared after 2–3 d. Crystal quality was increased by exchanging the mother liquor with saturated sodium formate solution for 24 h. Dehydrated crystals were frozen without the addition of any other cryo-protectant. Data sets were collected either at the Deutsche Elektronen Synchrotron (DESY, Hamburg) on beamline BW6 or at the Berliner Elektronenspeicher Synchrotron (BESSY, Berlin) on PSF beamline BL2.

Wild-type Cnx1G and Ser583Ala crystals diffracted to 1.6 Å and 1.45 Å, respectively. The crystals belonged to cubic space group *F*₄32, with cell dimensions of $a = b = c = 171.03$ Å and $\alpha = \beta = \gamma = 90^\circ$ (Cnx1G), and contained one molecule per asymmetric unit. For Ser583Ala, a slightly different cell, with dimensions $a = b = c = 171.09$ Å, was obtained. The structure of Cnx1G was determined by molecular replacement using one monomer of Protein Data Bank (PDB) accession code 108N as a search model. Molecular replacement was done with the program AMORE²⁶ and the structure was refined by manual model building in O²⁷ and refinement with REFMAC²⁸.

The initial maps after molecular replacement showed electron density for all atoms of MPT. On the basis of the electron density, we chose MPT derived from sulphite oxidase¹³ as an appropriate coordinate model, with the exception that we removed the molybdenum before introducing it into the refinement process. The structure of Ser583Ala was determined by using the Cnx1G model without MPT to prevent any bias. After introducing MPT, the additional density extension next to MPT was identified as AMP, which was then introduced into the model. As MPT and AMP were linked for the refinement, O3P and O3PH were removed from AMP to build the pyrophosphate bond. Copper was introduced after the fluorescence scan of the Cnx1G crystals and subsequent analysis of the inflection point and peak data sets.

Determination of MPT, MPT–AMP and Moco

MPT and MPT–AMP content was determined by HPLC formA analysis²⁹. *In vitro* conversion of Cnx1G-bound MPT into Moco was determined using the *nit-1* reconstitution assay²⁰. All reconstitutions were done with freshly desalted *nit-1* extract that had been passed twice through Nick gel-filtration columns (Amersham) equilibrated with *nit-1* buffer¹⁴. *In vitro* Moco synthesis was done in a total volume of 20 μ l of *nit-1* buffer containing 1 mM reduced glutathione, 10 μ l of *nit-1* protein extract and 1.5–6 pmol MPT–protein complex, according to the linear range of the assay. Other additives are indicated. The reaction was incubated for 1 h at room temperature, and the activity of the reconstituted NADPH–nitrate reductase was then determined¹⁴. Nitrite was detected at an absorbance of 540 nm using a Versa max 96-well plate reader (Molecular Devices). One unit of Moco activity was defined as the amount of reconstituted *nit-1* nitrate reductase sufficient to produce an increase at 540 nm of 1.0 absorbance units per 1 nmol Cnx1G in 25 min of reaction time.

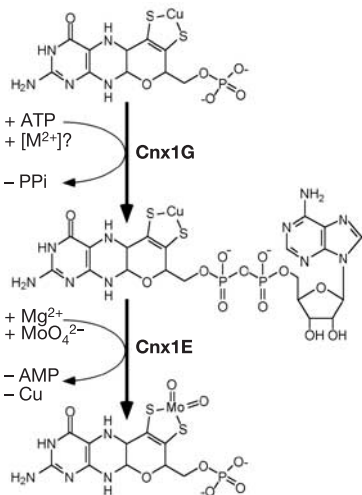


Figure 4 Hypothetical mechanism of the molybdenum insertion reaction. The mechanism of copper chelation is unknown. Cnx1G adenylates Cu–MPT, a reaction that is thought to be cation (M^{2+}) dependent. Cnx1E subsequently cleaves the Cu–MPT–AMP complex to activate and to transfer the molybdenum atom to MPT. The function of copper is seen in the protection of the reactive MPT dithiolate.

Data Bank accession numbers

The atomic coordinates and structure factors of Cnx1G (1uux) and Ser583Ala (1uuy) have been deposited in the PDB.

Received 11 March; accepted 18 May 2004; doi:10.1038/nature02681.

- Hille, R. The mononuclear molybdenum enzymes. *Chem. Rev.* **96**, 2757–2816 (1996).
- Romao, M. J. *et al.* Crystal structure of the xanthine oxidase-related aldehyde oxidoreductase from *D. gigas*. *Science* **270**, 1170–1176 (1995).
- Chan, M. K., Mukund, S., Kletzin, A., Adams, M. W. W. & Rees, D. C. Structure of a hyperthermophilic tungstopterin, aldehyde ferredoxin oxidoreductase. *Science* **267**, 1463–1469 (1995).
- Mendel, R. R. & Schwarz, G. Biosynthesis and molecular biology of the molybdenum cofactor (Moco). *Met. Ions Biol. Syst.* **39**, 317–368 (2002).
- Stiefel, E. I. Molybdenum bolsters the bioinorganic brigade. *Science* **272**, 1599–1600 (1996).
- Enemark, J. H. & Cosper, M. M. Molybdenum enzymes and sulfur metabolism. *Met. Ions Biol. Syst.* **39**, 621–654 (2002).
- Mendel, R. R. & Schwarz, G. Molybdoenzymes and molybdenum cofactor in plants. *Crit. Rev. Plant. Sci.* **18**, 33–69 (1999).
- Johnson, J. L. *et al.* Inborn errors of molybdenum metabolism: combined deficiencies of sulfite oxidase and xanthine dehydrogenase in a patient lacking the molybdenum cofactor. *Proc. Natl Acad. Sci. USA* **77**, 3715–3719 (1980).
- Shih, V. E. *et al.* Sulfite oxidase deficiency. Biochemical and clinical investigations of a hereditary metabolic disorder in sulfur metabolism. *N. Engl. J. Med.* **297**, 1022–1028 (1977).
- Kuper, J., Winking, J., Hecht, H. J., Mendel, R. R. & Schwarz, G. The active site of the molybdenum cofactor biosynthetic protein domain Cnx1G. *Arch. Biochem. Biophys.* **411**, 36–46 (2003).
- Schwarz, G. *et al.* The molybdenum cofactor biosynthetic protein Cnx1 complements molybdate-repairable mutants, transfers molybdenum to the metal binding pterin, and is associated with the cytoskeleton. *Plant Cell* **12**, 2455–2472 (2000).
- Kuper, J. *et al.* In vivo detection of molybdate-binding proteins using a competition assay with MoDE in *Escherichia coli*. *FEMS Microbiol. Lett.* **218**, 187–193 (2003).
- Kisker, C. *et al.* Molecular basis of sulfite oxidase deficiency from the structure of sulfite oxidase. *Cell* **91**, 973–983 (1997).
- Kuper, J., Palmer, T., Mendel, R. R. & Schwarz, G. Mutations in the molybdenum cofactor biosynthetic protein Cnx1G from *Arabidopsis thaliana* define functions for molybdopterin bind, Mo-insertion and molybdenum cofactor stabilization. *Proc. Natl Acad. Sci. USA* **97**, 6475–6480 (2000).
- Lake, M. W., Wuebbens, M. M., Rajagopalan, K. V. & Schindelin, H. Mechanism of ubiquitin activation revealed by the structure of a bacterial MoeB–MoaD complex. *Nature* **414**, 325–329 (2001).
- Schindelin, H., Kisker, C., Hilton, J., Rajagopalan, K. V. & Rees, D. C. Crystal structure of DMSO reductase: redox-linked changes in molybdopterin coordination. *Science* **272**, 1615–1621 (1996).
- Schrader, N. *et al.* The crystal structure of plant sulfite oxidase provides insights into sulfite oxidation in plants and animals. *Structure* **11**, 1251–1263 (2003).
- Santamaria-Araujo, J. A. *et al.* The tetrahydropyranopterin structure of the sulfur-free and metal-free molybdenum cofactor precursor. *J. Biol. Chem.* **279**, 15994–15999 (2004).
- Bertero, M. G. *et al.* Insights into the respiratory electron transfer pathway from the structure of nitrate reductase A. *Nature Struct. Biol.* **10**, 681–687 (2003).
- Nason, A. *et al.* In vitro formation of assimilatory reduced nicotinamide adenine dinucleotide phosphate: nitrate reductase from a *Neurospora* mutant and a component of molybdenum-enzymes. *Proc. Natl Acad. Sci. USA* **68**, 3242–3246 (1971).
- Xiang, S., Nichols, J., Rajagopalan, K. V. & Schindelin, H. The crystal structure of *Escherichia coli* MoeA and its relationship to the multifunctional protein Gephyrin. *Structure* **9**, 299–310 (2001).
- Schrag, J. D. *et al.* The crystal structure of *Escherichia coli* MoeA, a protein from the molybdopterin synthesis pathway. *J. Mol. Biol.* **310**, 419–431 (2001).
- Moorhead, G. B. *et al.* Purification of a plant nucleotide pyrophosphatase as a protein that interferes with nitrate reductase and glutamine synthetase assays. *Eur. J. Biochem.* **270**, 1356–1362 (2003).
- Mercer, J. F. The molecular basis of copper-transport diseases. *Trends Mol. Med.* **7**, 64–69 (2001).
- Mason, J. Thiomolybdates: mediators of molybdenum toxicity and enzyme inhibitors. *Toxicology* **42**, 99–109 (1986).
- Navaza, J. AMORE—an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 157–163 (1994).
- Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
- Murshudov, G., Vagin, A. & Dodson, E. Refinement of macromolecular structures by the maximum likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
- Guse, A. *et al.* Biochemical and structural analysis of the molybdenum cofactor biosynthesis protein MobaA. *J. Biol. Chem.* **278**, 25302–25307 (2003).
- Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct. Funct. Genet.* **11**, 281–296 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. N. Pau for inspiring discussion; T. Otte and F. Koenig for technical assistance; the staff at beamlines BW6 at DESY and PSF-BL2 at BESSY; and V. Wray for critically reading the manuscript. This work was supported by grants from the Deutsche Forschungsgemeinschaft (to H.J.H., R.R.M. and G.S.), and the Fonds der Chemischen Industrie and the Fritz Thyssen Stiftung (to R.R.M.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.S. (g.schwarz@tu-bs.de).

Electron microscopic analysis of KvAP voltage-dependent K⁺ channels in an open conformation

Qiu-Xing Jiang¹, Da-Neng Wang² & Roderick MacKinnon¹

¹Howard Hughes Medical Institute and Laboratory of Molecular Neurobiology and Biophysics, Rockefeller University, 1230 York Avenue, New York, New York 10021, USA

²Skirball Institute of Biomolecular Medicine, New York University Medical Center, 540 First Avenue, New York, New York 10016, USA

Voltage-dependent ion channels serve as field-effect transistors by opening a gate in response to membrane voltage changes¹. The gate's response to voltage is mediated by voltage sensors², which are arginine-containing structures that must move with respect to the membrane electric field. We have analysed by electron microscopy a voltage-dependent K⁺ channel from *Aeropyrum pernix* (KvAP)³. Fab fragments were attached to 'voltage sensor paddles' and identified in the electron microscopy map at 10.5 Å resolution. The extracellular surface location of the Fab fragments in the map is consistent with the membrane-depolarized, open conformation of the channel in electrophysiological experiments. Comparison of the map with a crystal structure⁴ demonstrates that the voltage sensor paddles are 'up' (that is, near the channel's extracellular surface) and situated at the protein–lipid interface. This finding supports the hypothesis that in response to changes in voltage the sensors move at the protein–lipid interface⁵ rather than in a gating pore surrounded by protein^{6,7}.

Voltage-dependent cation channels have voltage sensors that contain a hydrophobic α-helix with arginine residues known as S4^{1,2}. The movement of the cationic arginine residues relative to the membrane electric field energetically couples the pore conformation to transmembrane voltage^{1,2}. How does such a movement occur? The crystal structure of a voltage-dependent K⁺ channel KvAP showed that in each subunit S4 forms part of a helix-turn-helix structure termed a voltage sensor paddle⁴. In the crystal the paddles, which are attached to the pore by flexible hinges, are in an apparently non-native conformation (at the intracellular surface), probably owing to crystal packing constraints and the absence of a membrane. Despite uncertainties in the relationship between the KvAP crystal structure and its possible conformations in the membrane, the crystal structure conveyed a very interesting idea: that the voltage sensor paddles probably move relative to the pore for it to open. Experiments with biotin and avidin have lent strong support for this idea⁵. The voltage sensor paddles were hypothesized to move within the membrane at the protein–lipid interface. A more complete understanding of voltage-dependent gating, however, requires additional experiments, especially new protein structures under a variety of conditions.

We performed single particle reconstruction⁸ from electron microscopy images of purified KvAP protein to obtain a three dimensional (3D) structure of the channel. In this method the channel is not subject to crystal packing forces. We attached four 33H1 Fab fragments, one to each voltage sensor, with three intentions. First, the mass of one KvAP tetramer is ~100 kDa, which is too small for accurate cryo-electron microscopy analysis⁹. Four Fab fragments increase the total particle mass to ~300 kDa, large enough for accurate single particle alignment. Second, the 33H1 Fab fragment binds to the voltage sensor in electrophysiological experiments only when the channel opens, and only from the extracellular side of the membrane (Fig. 1a)⁵. Therefore, we can assess whether the Fab positions are consistent with the electrophysiological experiments. Third, a crystal structure of the 33H1

Fab bound to the voltage sensor defines (at 1.9 Å resolution) the precise location of the voltage sensor paddle, which is fixed with respect to the Fab fragment (Fig. 1b)⁴. Therefore, by identifying the Fab fragments in the map we can infer the locations of the voltage sensor paddles.

An initial model was generated from images of single particles stained with ammonium molybdate (Fig. 1c and see Methods). This initial model was used as a starting point for analysis of cryo-negatively stained images (Fig. 1d). The use of negative stain in vitrified ice has been shown to increase the contrast of single particles¹⁰, and limits the resolution¹¹ to the size of the stain molecules (~8.5 Å). Sufficient contrast allowed interactive particle picking and reliable alignment for single particle reconstruction (Fig. 1d). A 3D map was generated from ~21,400 particles (Fig. 1e). The resolution was estimated to be 10.5 Å based on a 0.5 threshold in the Fourier shell correlation (FSC, Fig. 1f)^{12,13}. As expected at this resolution, many surface features are easily discernable, including α-helices projecting from the channel surface (Fig. 2a), two identifiable immunoglobulin domains for each Fab (Fig. 1e) and a

groove separating the main β sheets within the variable immunoglobulin domain (Fig. 2e). Heterogeneity in the position of the constant domain with respect to the variable domain is evident in the map, as has been described in other electron microscope reconstructions with Fab fragments¹⁴. These features confirm that the map contains the structural details consistent with the estimation of resolution by FSC of 0.5, which has been shown to be a conservative definition of resolution¹⁵.

To constrain the voltage sensor paddle with respect to the pore we docked the crystal structures of the pore and the Fab–paddle into the map independently. The pore from the KvAP crystal structure⁴ fits well without modification, with the S6 (inner) helices, inserting into helical tubes projecting from the intracellular surface of the map (Fig. 2b). The recognizable chirality of this structure defines unambiguously the handedness of the map. On the extracellular side, the level of the selectivity filter opening and turret structures conform to the map's surface (Fig. 2c). The position of the pore is therefore accurately defined with respect to rotation and translation.

The constant immunoglobulin domain of Fab and the variable domain of Fab, coupled to the voltage sensor paddle, were docked separately. The constant domain was not very well constrained because of its apparent intrinsic mobility in the particles. But the variable domain, which is attached directly to the voltage sensor paddle, conforms very well both to the volume and surface features of the map, allowing its accurate placement (Fig. 2d, e), with a single ambiguity. A Fab fragment has a pseudo two-fold axis parallel to its longest dimension (Fig. 1b), and therefore at 10.5 Å resolution there are two (but only two) possible ways related by 180° rotation to place the Fab fragment into the map. These two ways define two possible orientations of the voltage sensor paddle with respect to the pore, because the paddle is fixed relative to the Fab fragment (Fig. 1b). These are shown, as viewed from the side (Fig. 3a, b) and in cross-section near the level of the extracellular surface (Fig. 3c, d).

The location of density for the Fab fragments at the channel's extracellular surface in the map (Figs 1e and 2c) is consistent with electrophysiological studies showing that 33H1 Fabs bind only from the extracellular side and only when a channel opens (Fig. 1a)⁵. This correlation between structure and activity of the Fabs in electrophysiological experiments leads us to conclude that the voltage sensors are in an open conformation.

In either of the two possible orientations, the voltage sensor paddles are located at the lateral (membrane-facing) edge of the map, at the channel's outer perimeter with an entire face of S4 exposed to what would be the membrane in the native state (Fig. 3a, b). 'Extra density' not accounted for by the partial model consisting of the pore, Fabs and voltage sensor paddles presumably contains the S1, S2 and S3a helices. This density is located mainly between the voltage sensor paddles from neighbouring subunits (red dashed ovals, Fig. 3c, d), not peripheral to the voltage sensor paddles. In several studies disulphide cross-links connecting S4 to the pore^{16–18} have been given as evidence against S4 being located at the protein–lipid interface, but Fig. 3 suggests that such cross-links and a perimeter location for S4 are not mutually exclusive.

The electron microscopy map places an additional constraint on the S4 helix beyond the X-ray structures. In the two possible orientations of the paddle, the long straight S4 helix from the X-ray structure⁴ of the isolated voltage sensor projects well beyond the surface (Fig. 4a, b), which implies that the helix must bend to fit the map. The red arrows (Fig. 4a, b) point to a glycine residue (Gly 134) where the S4 is bent in the full-length crystal structure of KvAP⁴; the correspondence of this glycine with the surface of the map implies that the S4 helix is also bent at this position in the single particle structure. In the KvAP crystal structure⁴ the S4 was defined as the hydrophobic, arginine-containing helix amino-terminal to Gly 134, forming one helix of the voltage sensor paddle, whereas the S4–S5 linker was defined as the amphipathic helical segment running from Gly 134 to the N terminus of S5 (Fig. 4c). The map requires this

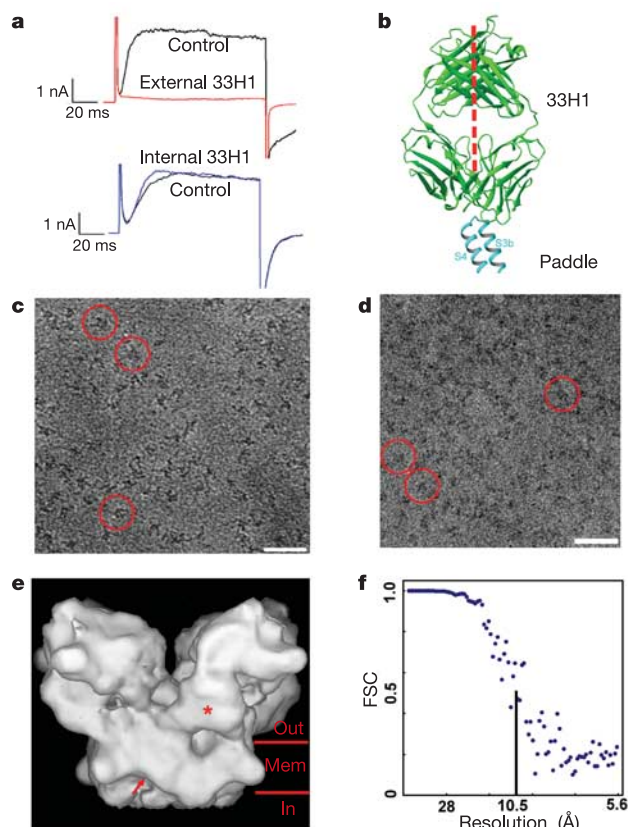


Figure 1 Structure of the KvAP–33H1 complex at 10.5 Å. **a**, Functional effect of the 33H1 Fab on the ionic current carried by KvAP channels in planar membranes. Depolarization to +100 mV elicited an outward K⁺ current (black traces). Fab 33H1 inhibited ionic current from the extracellular side (red trace), but not from the intracellular side (blue trace). **b**, Structure of the 33H1 Fab fragment bound to the voltage sensor paddle as determined by X-ray crystallography⁴ (PDB code 1ORS). The paddle is rigidly attached to the Fab fragment. The red dashed line represents the pseudo two-fold axis of the Fab fragment. **c**, Image of negatively stained KvAP–33H1 complexes. Protein appears black here and in **d**. Three particles are marked with red circles. The scale bar is 50 nm. **d**, Image of cryo-negatively stained KvAP–33H1 complexes. Select particles are marked with red circles. The particles were well separated and took different orientations. The scale bar is 50 nm. **e**, 3D map of the KvAP–33H1 complex at 10.5 Å. The red arrow identifies a position considered in Fig. 4a and b. The red asterisk marks the variable immunoglobulin domain and area shown in Fig. 2d. **f**, Fourier shell correlation (FSC) for the 3D map of the KvAP–33H1 complex from two maps calculated using two halves of the data. The 0.5 threshold corresponds to 10.5 Å.

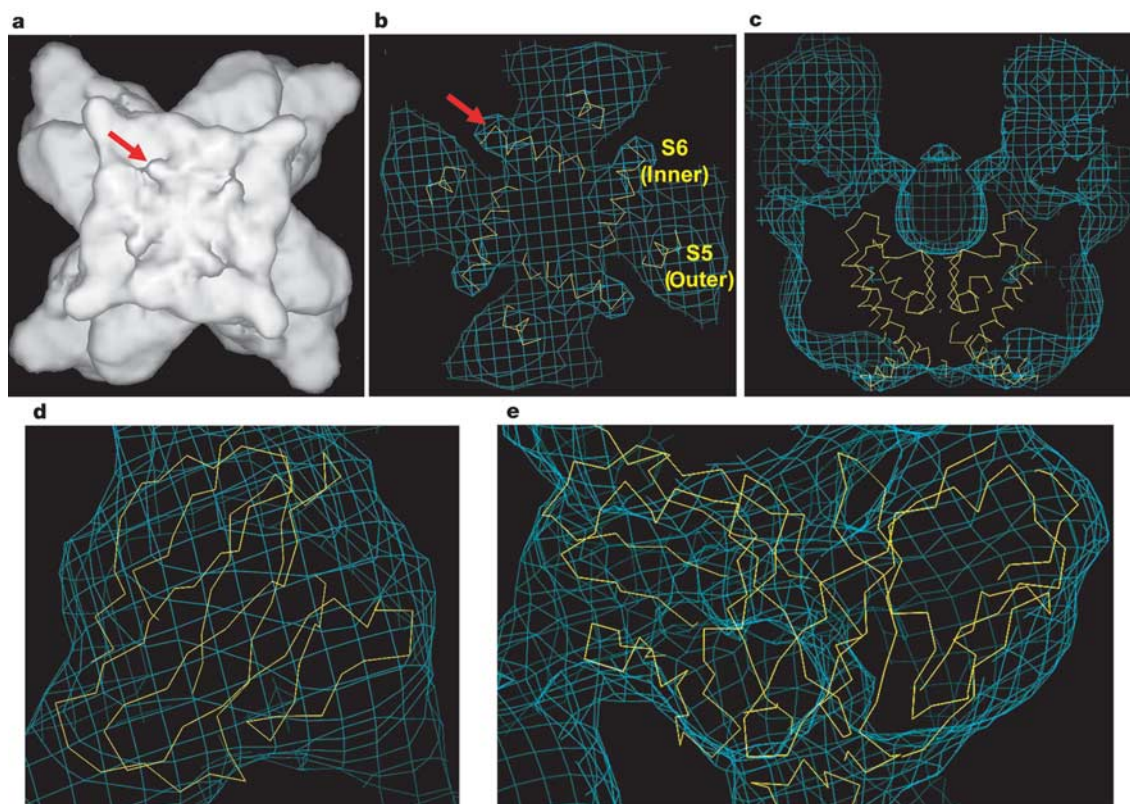


Figure 2 Docking of the crystal structures of the KvAP pore and 33H1Fab fragments into the electron microscope map. **a**, View of the map of the KvAP–33H1 complex from the intracellular side. The short tubes (red arrow) show α -helices projecting from the surface. **b**, Cross section near the intracellular surface of the map (blue mesh) with the KvAP pore (PDB code 1ORQ, yellow traces) in position. The ends of S6 correspond to the protruding

tubes (red arrow). S5 projects into the density beside S6. **c**, Side-view of a central section of the map (blue mesh) with the docked pore (yellow trace). **d**, **e**, Docking of the variable immunoglobulin domain of the Fab (PDB code 10RS, yellow trace) into the map (blue mesh). The view in **d** was identified by a red asterisk in Fig. 1e. The map in **e** was rotated 90° anti-clockwise along the vertical axis from the orientation in **d**.

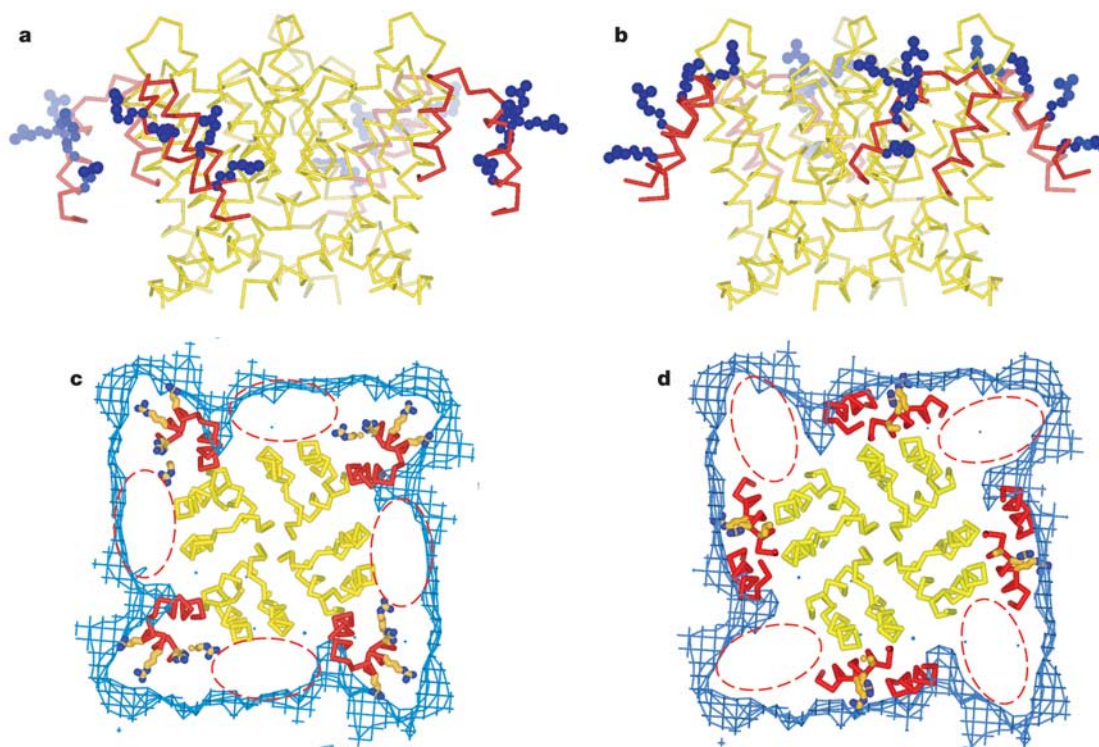


Figure 3 Two possible orientations of the paddle relative to the pore. **a**, **b**, Two orientations of the paddle (red) are shown beside the pore (yellow). Arginine residues on S4 are shown in blue. **c**, **d**, Cross sections of the map (blue mesh) at the extracellular

surface with the paddle and the pore in position. The extra density (highlighted by red dashed ovals) could be accounted for by S1, S2 and S3a for each possible orientation of the paddle.

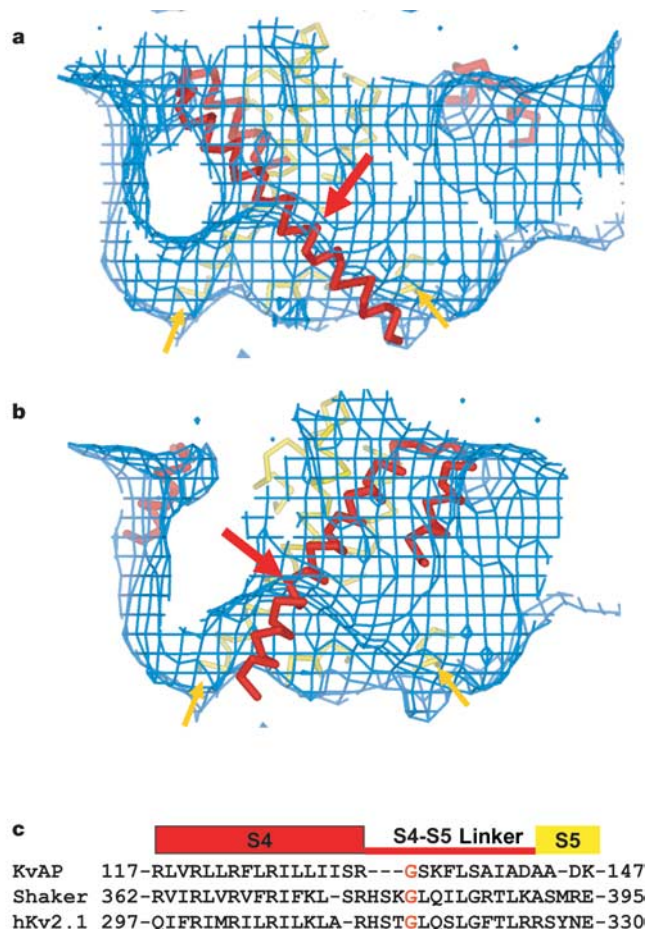


Figure 4 The S4 probably bends and becomes the S4–S5 linker at Gly 134. **a, b**, The long S4 helix (red helix) from the crystal structure of the isolated voltage sensor (PDB code 1ORS) projects beyond the map (blue mesh). In either possible orientation, Gly 134 (red arrow) is located near the surface of the map. The S4 is bent at this position in the crystal structure of full-length KvAP (PDB code 1ORQ). Yellow arrows show the locations of nearby S5 helices. **c**, Alignment of S4 and S4–S5 linker sequences of KvAP, *Shaker* and human Kv2.1 potassium channels. The linker sequence is not conserved, but has a distinct amphipathic feature with a glycine residue (highlighted in red) close to its beginning.

linker to run parallel to the membrane surface to reach a nearby S5 helix, although we can not tell from the map which adjacent S5 it is attached to (yellow arrows, Fig. 4a, b). This suggests that the S4 helix is only about five helical turns, not much longer than the length of the paddle, and that the S4–S5 linker (originally proposed to be part of S4 based on sequence alignments)¹⁹ is interfacial near the intracellular solution. These assignments of S4 and linker are consistent with functional data showing that the first four arginine residues in the *Shaker* K⁺ channel of *Drosophila* carry most of the gating charge^{20,21}.

The electron microscopy data give us a picture of a voltage-dependent K⁺ channel with its voltage sensors in the membrane-depolarized, open conformation. In contrast to the crystal structure of KvAP in which the voltage sensor paddles adopted an intracellular position (but not a closed position)⁴, the voltage sensor paddles in the single particles are 'up'. There are two very obvious and important aspects to the electron microscopy structure. First, the voltage sensor paddles are very near the extracellular solution. If one were to imagine a membrane surrounding the channel then the four highly conserved S4 arginine residues would be electrically near the extracellular solution. In other words, the gating charges would be displaced to their open channel positions. Second, the voltage sensor paddles are at the perimeter of the channel. This position

of the paddles at the perimeter is not consistent with the conventional model in which S4 is hypothesized to rotate or translate within a protein-lined, aqueous gating pore^{6,7}. It is consistent with the hypothesis that the voltage sensor moves at the protein–lipid interface when the channel undergoes a conformational change from the closed to open state^{4,5}. □

Methods

Preparation of KvAP–33H1 complexes

The KvAP channel protein was expressed in XL1-blue strain *Escherichia coli* and purified as described previously³. The 33H1 Fab was prepared according to standard protocols⁴. Stable KvAP–33H1 complexes were purified on a Superdex 200 size-exclusion column. The protein was concentrated to about 5.0 mg ml^{−1} with a Millipore concentrator (100 kDa molecular weight cut off) in a buffer containing 5 mM n-Decyl-β-D-Maltopyranoside, 40 mM KCl, 60 mM NaCl and 20 mM Tris-HCl at pH 7.4.

Negative stain and cryo-negative stain electron microscopy

For negative staining, freshly carbon-coated copper grids were glow-discharged for 2 min immediately before use. The KvAP–33H1 complexes of 0.10 mg ml^{−1} were applied to the grid. After 1 min the solution was blotted almost completely away, to leave a thin layer of sample on the surface. The grid was then quickly floated on top of a drop of stain solution (6.0% ammonium molybdate-NaOH, pH 6.7–7.4 and 2.0% trehalose). After 10–15 s, the stain solution was completely blotted, and the grid was dried in a desiccator under vacuum for 3 h.

To prepare cryo-negative stain samples, holey grids were covered with a layer of thin carbon film (~15 nm) and glow-discharged for 2.0 min before use. Immediately before freezing, a saturated solution of ammonium molybdate (~65%, pH adjusted to 6.7–7.4 with 10.0 M NaOH) was prepared. Equal volumes of the stain solution and the protein sample (~5.0 mg ml^{−1}) were mixed and 3.0 μl of the mixture was applied to a holey grid. After about 10 s the grid was transferred to a guillotine plunger, blotted with a slip of filter paper, and immediately plunged into liquid ethane. The cryo-specimens were stored in liquid nitrogen until electron microscope examination.

The negative-stain specimens were examined in a Phillips CM12 microscope operated at 120 kV, and the cryo-negative stain specimens in either the CM12 or a Phillips CM200 FEG microscope run at 200 kV. The defocus range was −0.5 to −1.1 μm. The low-dose mode was used to take pictures using Kodak SO-163 films at × 60,000 magnification with the CM12 and × 50,000 magnification with the CM200. For cryo-negative stain specimens the electron dose was kept at ~20 e₀ per Å² per exposure.

Image analysis

Films were developed in full-strength D19 for 12 min. Images were examined on an optical diffractometer and only those with no obvious drift or astigmatism were digitized in a Zeiss SCAI scanner (ZI imaging) at 14 μm steps, corresponding to 2.8 Å on the specimen level with the CM200 microscope.

The analysis of the images followed standard protocols^{22,23}. The defocus level for each electron micrograph was determined by CTF fitting using a MATLAB program (the Mathworks). The B factors estimated from the CTF fitting of the films were ~100 Å². Single particles were interactively selected from each image by use of the BOXER program in the EMAN package²⁴. The particle images were extracted and then phase-flipped with the CTFIT in EMAN. The IMAGIC package²⁵ was used to generate two initial models by angular reconstitution²⁶ from a negative-stain data set (~4,500 particle images) and a cryo-negative stain data set (~8,200) collected in the CM12. The two models were at 28 Å and 19 Å resolution, respectively. Multivariate statistical analysis of both data sets showed strong four-fold symmetry, consistent with the channels being tetramers with four Fab fragments bound to each channel. These two models were taken as the references to process in the SPIDER package²⁷ a large data set of 28,960 particles picked from 36 micrographs taken in the CM200 FEG microscope. The two initial models gave rise to the same final map after the refinement. 90% of the particles were selected based on the correlation coefficient with the reference map. The final map was calculated from 21,379 particles after overpopulated classes were limited in size in the final reconstruction. For amplitude correction, a B factor of 100 Å² and an amplitude contrast constant of 15% (ref. 28) were used in the Wiener filtering²⁹, which had little effect on the map. Data assessing the accuracy with which the final map represents the particle images are presented in Supplementary Information. The resolution of the final map was estimated by applying the 0.5 threshold to the Fourier shell correlation^{12,13} between two maps calculated from two halves of the data. The map was contoured at a density level to give ~340 kDa to the KvAP–Fab complex, taking into account about 40 kDa for the detergent micelle.

The docking of the crystal structures of the KvAP pore (protein data bank (PDB) code 1ORQ) and the Fab fragment–voltage sensor paddle (PDB code 1ORS) was performed manually using the program O (ref. 30).

Received 17 March; accepted 3 June 2004; doi:10.1038/nature02735.

1. Sigworth, F. Voltage gating of ion channels. *Q. Rev. Biophys.* **27**, 1–40 (1994).
2. Bezanilla, F. The voltage sensor in voltage-dependent ion channels. *Physiol. Rev.* **80**, 555–592 (2000).
3. Ruta, V., Jiang, Y., Lee, A., Chen, J. & MacKinnon, R. Functional analysis of an archaebacterial voltage-dependent K⁺ channel. *Nature* **422**, 180–185 (2003).
4. Jiang, Y. *et al.* X-ray structure of a voltage-dependent K⁺ channel. *Nature* **423**, 33–41 (2003).
5. Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. The principle of gating charge movement in a voltage-dependent K⁺ channel. *Nature* **423**, 42–48 (2003).
6. Bezanilla, F. Voltage sensor movements. *J. Gen. Physiol.* **120**, 465–473 (2002).

7. Horn, R. Coupled movements in voltage-gated ion channels. *J. Gen. Physiol.* **120**, 449–453 (2002).
8. Frank, J. Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 303–319 (2002).
9. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28**, 171–193 (1995).
10. Adrian, M., Dubochet, J., Fuller, S. D. & Harris, J. R. Cryo-negative staining. *Micron* **29**, 145–160 (1998).
11. Golas, M. M., Sander, B., Will, C. L., Luhrmann, R. & Stark, H. Molecular architecture of the multiprotein splicing factor SF3b. *Science* **300**, 980–984 (2003).
12. Saxton, W. O. & Baumeister, W. The correlation averaging of a regularly arranged bacterial cell envelope protein. *J. Microsc.* **127**, 127–138 (1982).
13. van Heel, M. Similarity measures between images. *Ultramicroscopy* **21**, 95–100 (1987).
14. Conway, J. F. *et al.* Characterization of a conformational epitope on hepatitis B virus core antigen and quasiequivalent variations in antibody binding. *J. Virol.* **77**, 6466–6473 (2003).
15. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333**, 721–745 (2003).
16. Laine, M. *et al.* Atomic proximity between S4 segment and pore domain in *Shaker* potassium channels. *Neuron* **39**, 467–481 (2003).
17. Gandhi, C. S., Clark, E., Loots, E., Pralle, A. & Isacoff, E. Y. The orientation and molecular movement of a K⁺ channel voltage-sensing domain. *Neuron* **40**, 515–525 (2003).
18. Neale, E. J., Elliott, D. J., Hunter, M. & Sivaprasadarao, A. Evidence for intersubunit interactions between S4 and S5 transmembrane segments of the *Shaker* potassium channel. *J. Biol. Chem.* **278**, 29079–29085 (2003).
19. Chandry, K. G. & Gutman, G. A. In *Ligand and Voltage-Gated Channels* (ed. North, R. A.) 1–72 (CRC, Boca Raton, 1995).
20. Seoh, S. A., Sigg, D., Papazian, D. M. & Bezanilla, F. Voltage-sensing residues in the S2 and S4 segments of the *Shaker* K⁺ channel. *Neuron* **16**, 1159–1167 (1996).
21. Aggarwal, S. K. & MacKinnon, R. Contribution of the S4 segment to gating charge in the *Shaker* K⁺ channel. *Neuron* **16**, 1169–1177 (1996).
22. Orlova, E. V. *et al.* Structure of keyhole limpet hemocyanin type 1 (KLH1) at 15 Å resolution by electron cryomicroscopy and angular reconstitution. *J. Mol. Biol.* **271**, 417–437 (1997).
23. Gabashvili, I. S. *et al.* Solution structure of the *E. coli* 70S ribosome at 11.5 Å resolution. *Cell* **100**, 537–549 (2000).
24. Ludtke, S. J., Baldwin, P. R. & Chiu, W. EMAN: semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* **128**, 82–97 (1999).
25. van Heel, M., Harauz, G., Orlova, E. V., Schmidt, R. & Schatz, M. A new generation of the IMAGIC image processing system. *J. Struct. Biol.* **116**, 17–24 (1996).
26. van Heel, M. Angular reconstitution: a posteriori assignment of projection directions for 3D reconstruction. *Ultramicroscopy* **21**, 111–123 (1987).
27. Frank, J. *et al.* SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. *J. Struct. Biol.* **116**, 190–199 (1996).
28. Sander, B., Golas, M. M. & Stark, H. Automatic CTF correction for single particles based upon multivariate statistical analysis of individual power spectra. *J. Struct. Biol.* **142**, 392–401 (2003).
29. Hawkes, P. W. In *Computer Processing of Electron Microscopic Images* (ed. Hawkes, P. W.) 1–33 (Springer, Berlin, 1980).
30. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank members of the MacKinnon lab, S. Darst, N. Opalka and D. Stokes for helpful discussions. This work was supported by grants from the NIH to D.N.W. and R.M. R.M. is an investigator in the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M. (mackinn@rockefeller.edu). The electron microscopy map has been deposited with European Bioinformatics Institute (www.ebi.ac.uk/msd/emd-2771.map).

corrigenda

Enzymic activation and transfer of fatty acids and acyl-adenylates in mycobacteria

Omita A. Trivedi, Pooja Arora, Vijayalakshmi Sridharan, Rashmi Tickoo, Debasisa Mohanty & Rajesh S. Gokhale

Nature **428**, 441–445 (2004).

The structures of lauroyl-AMP and lauroyl-CoA are incorrectly represented in Fig. 1c, d of this Letter, which shows the L-form rather than the natural D-form of adenosine; there is no phosphate group at the 2'-position of the ribosyl moiety of lauroyl-AMP (Fig. 1c) and the phosphate group in lauroyl-CoA (Fig. 1d) should be at the 3'-position. These errors do not affect our conclusions. □

Reduction of hysteresis losses in the magnetic refrigerant Gd₅Ge₂Si₂ by the addition of iron

Virgil Provenzano, Alexander J. Shapiro & Robert D. Shull

Nature **429**, 853–857 (2004).

It has come to our attention that we failed to cite in our Letter to *Nature* a key publication¹ that describes changes in the magnetocaloric capabilities of Gd₅(Si₂Ge₂) as a result of alloying the compound with small amounts of various metals, including iron. That paper supports the findings presented in our Letter. However, we also show that it is important to consider the magnetic hysteresis in addition to the magnetocaloric effect when assessing the usefulness of a material as a magnetic refrigerant. □

1. Pecharsky, V. K. & Gschneidner, K. A. Effect of alloying on the giant magnetocaloric effect of Gd₅(Si₂Ge₂). *J. Magn. Magn. Mater. Res.* **167**, L179–L184 (1997).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:

Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Scandinavia/ Spain/ Portugal:

Evelina Rubio Håkansson (4973)

Natureevents: Silke Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Stefan Hales

European Satellite Office

Germany/ Austria/ Italy/

The Netherlands/ Belgium:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami
e-mail: rasami@naturejp.com

naturejobs

Making the switch

The sad death of Francis Crick at the end of last month rightly generated a number of tributes from his friends and colleagues. But as I read these words, I was struck by a surprising fact. Crick, the man who co-discovered the structure of DNA, did not begin work on his PhD until he was 35.

By today's standards, this is remarkably old. In France, if you have not achieved tenure by that age, your scientific career is considered by many to be over. In Germany, *Habilitation* — a process that requires PhD scientists to complete a second thesis and demonstrate their teaching abilities — means that many academics don't gain tenure until their forties, although reforms may change that. And in the United States, beginning a PhD at 35 would put you on track for tenure at the age of about 50.

Crick's career was slightly unusual — he had stints in the military and as an engineer before moving to the University of Cambridge. But his delayed entry into postgraduate studies should reassure others who would like to enter science later than is conventional. And it also points to ways in which 'late-bloomers' might reinvigorate the scientific workforce.

Encouraging people to begin a PhD after working in a different discipline or sector would create scientists who are more truly multidisciplinary. It would also provide an alternative pathway to the usual undergraduate–postgraduate–postdoc route that some young scientists find constraining. And targeted schemes to lure late-bloomers into labs may be one way to fill skill gaps that arise in a younger workforce that lacks training, education and experience.

Whatever the reason, it seems sensible to welcome skilled people who want to switch directions, rather than to impose barriers on them — and their potential breakthroughs.

Paul Smaglik
Naturejobs editor



Contents

POSTDOCS AND STUDENTS

Informal routes into
employment p812

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS

Networking, Networking, Networking, Networking, Networking

Drinks with a few dozen friends or a visit to an interesting employer: Myrna Watanabe meets groups finding informal ways into work.

Grace Wong thrusts a microphone into the face of a young woman in a crowded room. “Make a smart pitch,” Wong demands. The young woman has one minute to tell the audience who she is. Like them, she is taking part in a Nobel Pauling Biotech Symposium, named after the laureate who was one of Wong’s mentors and organized by Student Vision, a group Wong set up to aid biotechnology students. Besides supplying all the free dim sum participants can eat, the free symposia teach scientists of all age and skill levels how to network. They are not only put on the spot, but provided with formal talks, informal panels and plenty of schmoozing opportunities.

Scientists of all ages are coming up with informal clubs and functions — beyond the usual job fairs, meetings and graduate-student seminars — at which to meet potential employers or collaborators. They may have speakers from local biotech or drug companies, or sponsor ‘treks’ to nearby employers. Some members discover career options that they didn’t know existed.

LET’S GET TOGETHER

Most meetings include networking time, and that’s the main function of Biotech Tuesday in Boston. Some 160 people attend each monthly event, which frequently has several recruiters showing up, says Seth Taylor, managing general partner of life-sciences consultancy Vectur, who co-founded Biotech Tuesday in 2002. The success of these informal events has inspired new groups such as BioConnect in Connecticut. Sponsored by biomedical group Connecticut United for Research Excellence and Yale University’s Biotechnology Student Interest Group (SIG), BioConnect’s first event, in May, attracted at least two chief executives who were looking to hire people.

Bringing students and industry people together is a prime function of the clubs. The California Institute of



Red for go: offers on the table for the best and brightest at a job fair.

HUREWITZ CREATIVE/CORBIS

GSAS HARVARD BIOTECH CLUB



Jiahong Juda sees networking events as a chance to develop skills.

Technology is located in Pasadena, well away from the California critical masses of San Diego and the San Francisco Bay Area. The closest biotech company is Amgen, about 68 kilometres away in Thousand Oaks. The Caltech Biotechnology Club has taken advantage of this proximity by having an 'Amgen series', where executives spoke about different aspects of the company's business. After the series concluded, the club sponsored a trip there and is now planning a similar jaunt to San Diego.

Graduate students and postdocs at the University of Pennsylvania in Philadelphia found that their science departments emphasized training for academia, so they organized the Penn Biotechnology Group to find out about opportunities elsewhere. Career treks have shown 600 or so members the scope of work available at nearby biotech and drug companies. Mohita Mohan, a biochemistry and molecular-biophysics graduate student, helped to organize a recent career trek to companies in King of Prussia, a suburb of Philadelphia. She found a large concentration of biotech firms there, but "they never tell you about this in your classes". On the trek, Mohan learned not to think that a biotech firm will automatically be hiring. Some are changing their model and not taking on new R&D personnel, and others, hoping to be acquired, will not necessarily offer secure employment.

Lectures on law, consulting firms, the business end of pharma and biotech, and venture-capital firms have also influenced members of the Penn group. "We had a few lectures on consulting — I didn't know they hired people with PhDs," says Cassia Cearley, a second-year graduate student in a neuroscience programme. She is now considering a career in consulting.

ENLIGHTENING LUNCH

Yale's SIG sponsors several different types of talk, including a formal lecture followed by lunch with the speaker and an informal discussion group with an invited guest. These have included a chief executive from a local biotech company, several corporate vice-presidents and someone from bioinformatics.

One discussion group featured a headhunter who talked about preparing for a job interview. Eric Anderson, a co-president of the Yale SIG, who recently completed a postdoc and is about to begin an MBA, learned through the SIG that there was more to science

than academics and cell biology. "It's given me a broader perspective of roles one can play," he says.

Faculty members and students at Johns Hopkins University in Baltimore, Maryland, joined to form the Hopkins Biotech Network, which has sponsored formal seminars. Co-founder and master's student Aditya Polsani says members appreciate the chance to meet industry people and gain a broader perspective on the industry.

Several clubs boast of informal networking leading to job offers. Anderson speaks of a Yale postdoc hired by a chief executive of New York City biotech Imclone Systems after a Yale SIG event. Taylor recalls a woman who got a speaking engagement at a prominent Harvard lab after attending a Biotech Tuesday meeting in Boston.

Some organizations are focused on more specific demographics. For instance, Women Entrepreneurs in Science and Technology (WEST) in Arlington Heights, Massachusetts, is for women who are already in the job market but wish to network for connections, skill development and job openings. The group's formal meetings and networking events help women learn how to use the entrepreneurial skills they have gained intellectually but have not used for career success, says Jiahong Juda, WEST's president and chief executive.

LEARNING FROM THE LEADERS

From the other side, industry organizations are also inviting young scientists to meet company leaders. In Britain, for example, the London Biotechnology Network sponsors monthly BioWednesdays which include networking. Although these are mainly meant for the industry, graduate students and postdocs are welcome. "We positively encourage young people to come along and see how businesses were started," says Marjory Goodall, healthcare executive with the economic development initiative London First, which founded the network in 2000.

In Australia, the industry body AusBiotech encourages students and postdocs to get involved. "Networking is the biggest factor attracting people to attend an event, so we always schedule time for this," says Fiona Smith of the Queensland branch. The branch holds five breakfast meetings a year at which three speakers may address a 'hot topic'. An annual careers night, originally for students, is now also aimed at people in the industry considering a career-path change. At September's event, six speakers will discuss their careers, to be followed by a showcase at which companies can meet potential employees.

About 250 people went to the New York Biotechnology Association's job fair in May, and 500 attended a day-long career seminar co-sponsored by the New York Academy of Sciences and *Naturejobs*. This featured separate panels for academia, biotech and industry. The EuroScience Open Forum 2004, to be held in Stockholm, Sweden, on 25–28 August, will have formal talks on careers sponsored by *Naturejobs*.

Wong, who doubles as president of ActoKine Therapeutics of Chestnut Hill, Massachusetts, is firmly behind the informal approach. She says that more than 20 people have received biotech or drug-company offers through the Nobel Pauling Biotech Symposia, proof that free dim sum and a microphone in the face can force interactions — and lead to jobs. ■

Myrna Watanabe is a freelance science writer in Patterson, New York.

Web links

Yale Biotechnology Student Interest Group

♦ www.yale.edu/biotech

Biotech Tuesday

♦ www.biotechtuesday.com

BioConnect

♦ www.curen.net

Women Entrepreneurs in Science & Technology

♦ WESTorg.org

Student Vision

♦ www.studentvision.org

Hopkins Biotech Network

♦ www.hopkinsbiotechnetwork.org

Caltech Biotechnology Club

♦ www.its.caltech.edu/~biotech

London Biotechnology Network

♦ www.londonbiotechnology.co.uk

AusBiotech

♦ www.ausbiotech.org

New York Biotechnology Association

♦ www.nyba.org